UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Propensity Scores: An Application in Interventional Cardiology

Luís António Garcez Duarte da Costa Ferreira

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Professora Doutora Maria Helena Mouriño Silva Nunes
Professora Doutora Ana Luisa Trigoso Papoila da Silva

2016

”I’ve missed more than 9000 shots in my career,
I’ve lost almost 300 games,
26 times, I’ve been trusted to take the game winning shot and missed.
I’ve failed over and over and over again in my life.
And that is why I succeed.”

Michael Jordan

Agradecimentos

Todas as pessoas que passam pela nossa vida nos ensinam algo. Desde já, deixo o meu expresso agradecimento a algumas destas pessoas que tornaram possível a escrita desta tese.

Ao Doutor Luis Raposo por me ter confiado os dados, permitindo assim a sua a análise e consequente escrita desta tese.

À Professora Doutora Helena Mouriño pela total disponibilidade em ouvir-me e partilhar opiniões, pelos conselhos e críticas, pelo empenho, pela paciência e pela exigência que sempre incutiu ao longo de todas as fases deste trabalho.

À Professora Doutora Ana Luisa Papoila pela constante motivação em todas as fases deste processo, por todos os conselhos e pela constante estimulação à produção de ideias inovadoras que deram origem à complexidade e profundidade desta tese.

Aos Professores do Mestrado em Bioestatística do DEIO, sem excepção, pela disponibilidade, abertura e por me terem proporcionado a melhor experiência de aprendizagem que eu me recordo. Ainda, um obrigado especial à Professora Doutora Lisete Sousa pelo acompanhamento e dedicação demonstrada para com todos os alunos.

À Faculdade de Ciências da Universidade de Lisboa por ter sido a minha 2ª casa durante 5 anos, e de onde guardo magníficas recordações.

A todos os meus amigos, pois tenho a sorte de ter muitos e bons que me acompanharam ao longo destes anos. Só com muito companheirismo, amizade, imaturidade, brincadeira e muitos jogos de LoL é possível sentir-se realizado no percurso académico.

À Inês por todo o carinho e paciência que teve comigo durante estes meses, pois foi essencial.

Aos meus Avós por todo o amor, aprendizagem, exemplo e carinho. Foram essenciais para que me tenha tornado na pessoa que sou hoje.

À minha Mãe e ao meu Pai, pela colaboração incondicional e confiança em mim. São os principais responsáveis por tudo isto.

Abstract

Invasive techniques are essential in investigation and management of cardiac and vascular diseases, particularly obstructive coronary artery disease. These invasive techniques can be performed for diagnostic or interventional purposes, and the access to the target arteries can be done through the femoral or the radial artery. The transradial approach may be clinically preferable, namely because its use has been associated with fewer peri-procedural complications, like bleeding. Despite the advantages over transfemoral approach, concerns have risen on the potential for transradial approach to increase the incidence of peri-procedural neurological complications, like Stroke or Transient Ischemic Attack (TIA) (Jurga et al., 2011). The aim of this work is to access the association between the risk of peri-procedural Stroke/TIA and the transradial or the transfemoral approaches.

A propensity score analysis was performed in a sample of 16 710 patients included in a single prospective registry between January of 2006 and November of 2012. Various PS methods like matching, genetic matching, stratification and inverse probability of treatment weighting were used to estimate the Average Treatment Effect for the Treated (ATT) and the Average Treatment Effect (ATE). To find the best possible matching, PS estimates were fitted from a GAMs and from logistic regressions. A logistic regression (LR) was performed too in order to identify all the factors associated to the occurrence of peri-procedural Stroke/TIA and estimate the Odds Ratios. To handle the low number of events and the subsequent separation problem, a Firth's logistic regression correction was run.

Both Propensity Score analysis and regression methods reached the same conclusions. The use of Transradial or Transfemoral Approach does not impact the occurrence of peri-procedural Stroke/TIA. So the clinically preferential use of Transradial can be performed without concerns relative to this technique. Regarding the methodology, GAM PS estimates provide better matchings when there are variables non-linear related with the treatment assignment logit, but genetic matching can overcome these differences by matching individually on these variables through the Generalized Mahalanobis Distance.

Keywords: Propensity Score, Generalized Additive Models, Genetic Matching, Stratification, Inverse Probability of Treatment Weighting

Resumo

A doença coronária é uma das principais causas de morbilidade e mortalidade à escala global quer nos no mundo moderno quer nos países em desenvolvimento devido ao estilo de vida das populações e ao progressivo envelhecimento da população. Os principais fatores de risco são a hipertensão arterial, a dislipidémia, a diabetes, o tabagismo e o sedentarismo (Sacco et al., 1997; Donnan et al., 2008). As manifestações clínicas dependem largamente do território vascular afectado.

Uma variedade de técnicas invasivas é utilizada em cardiologia, nomeadamente para o diagnóstico e tratamento e da doença coronária obstrutiva. Por norma, estes procedimentos usam um catéter (essencialmente um "tubo" inserido através de um ponto de acesso e guiado ao local de diagnóstico/intervenção, de modo a executar o procedimento). O ponto de acesso pelo qual o catéter é inserido é em geral a artéria femoral ou a artéria radial. Uma das complicações mais temidas da manipulação endovascular associada a este tipo de procedimentos é o Acidente Vascular Cerebral (AVC). A obstrução das artérias intra-craneanas devido a fenómenos trombo-embólicos provoca isquémia cerebral e em última análise, morte celular com perda temporária e definitiva de função neutológica (por exemplo, dificuldade na fala, dificuldade em compreender outras pessoas, perda de visão e falta de capacidade em sentir e mover certas partes do corpo) (Donnan et al., 2008).

Durante grande parte do século XX, o acesso transfemoral foi o mais utilzado pela maior parte dos cardiologistas de intervenção. Recentemente, tem sido demonstrado que o acesso transradial se associa a menos complicações peri-procedimento, como por exemplo a hemorragia, menos morbilidade e mortalidade, e tem menos custos que o acesso femoral. Todas estas vantagens contribuíram para a grande expansão do uso da artéria radial como ponto de acesso (Nathan and Rao, 2012; Burzotta et al., 2013).

No entanto, surgiram preocupações relativamente ao acesso transradial e ao risco de complicações neurológicas, sobretudo devido à maior manipulação do arco aórtico e das artérias subclávias (Jurga et al., 2011). O grande objetivo deste trabalho é estudar a associação entre a ocorrência de complicações neurológicas e o tipo de acesso usado durante o procedimento.

Os dados usados neste estudo observacional foram recolhidos a partir de cateterismos diagnósticos e de intervenção realizadas no Hospital de Santa Cruz, Carnaxide, entre Janeiro de 2006 e Novembro de 2012, sendo que os indivíduos incluídos no estudo têm 18 anos de idade, ou mais, tendo sido sujeitos a manipulação da artéria aorta ascendente ou do arco aórtico. A população final do estudo compreende 16710 indíviduos.

A primeira abordagem executada para estudar esta associação foi a regressão logística pois é a abordagem mais usada e considerada "standard". Para uma correta aplicação desta metodologia é conveniente ter 10 eventos por cada variável registada, para que assim os resultados sejam fiáveis (Peduzzi et al., 1996). No entanto, neste estudo apenas temos 27 eventos para 25 variáveis, o que põe em causa a

utilização da regressão logística. Este número baixo de eventos origina o problema da Separação, no qual não é possível calcular a real contribuição das variáveis "Transplante Renal" e "Histórico de AVC/AIT" na ocorrência do desfecho. Para contornar este problema, optou-se que pela aplicação da regressão logística de Firth, em que a estimação de máxima verosimilhança tem por base uma verosimilhança penalizada (Firth, 1993).

O baixo número de eventos e o grande número de variáveis a controlar (como é o caso) foi a principal motivação para a utilização da metodologia assente em Propensity Scores, além de outras vantagens referidas neste trabalho. O Propensity Score (PS) de cada individuo é, neste caso, a sua probabilidade de ter sido intervencionado pelo acesso transradial, sabendo determinado conjunto de variáveis. Tradicionalmente, os PS são calculados usando uma regressão logística. O cálculo destes pela regressão logística pressupõe que existe uma relação linear entre o logit da variável dependente (Acesso Transradial) e as covariáveis, pelo que, se este pressuposto não for verdadeiro, as estimativas dos PS estão erradas. Neste estudo, os PS também vão ser calculados via modelos aditivos generalizados (GAMs), pois este tipo de modelos são mais flexíveis e, portanto, conseguem descrever mais eficazmente a real relação entre uma covariável e a variável dependente, originando estimativas dos PS mais fiáveis (Woo et al., 2008).

A primeira metodologia de Propensity Scores usada é o "Matching", a qual consiste em emparelhar indivíduos que foram intervencionados pelo acesso transradial com outros intervencionados pelo acesso transfemoral. O "Matching" emparelha índividuos com PS semelhantes. Isto porque, está provado que dois indivíduos são assintoticamente comparáveis (mesma distribuição de variáveis) se tiverem PS estimados semelhantes (Rubin and Rosenbaum, 1983). Vários emparelhamentos foram executados, diferindo na forma de serem estimados os PS (usando a regressão logística ou modelos aditivos generalizados), ou diferindo nos conjuntos de variáveis usados como covariáveis nos modelos. Além do "Matching" simples, também foi executado o "Matching" genético, que utiliza a distância de Mahalanobis (distância multivariada) como medida de emparelhamento. Esta distância é calculada a partir das variáveis observadas durante o estudo e dos PS estimados. No "Matching" genético também foram experimentados vários conjuntos de variáveis, e os dois métodos de regressão para estimar PS. No final, o melhor "Matching" conseguido (de entre os "Matchings" simples e genéticos) foi um emparelhamento genético o qual incluiu todas as variáveis observadas e PS estimados a partir de um GAM. A qualidade do emparelhamento é medida através do quão homogéneas são as distribuições das variáveis nos dois grupos a comparar. O emparelhamento escolhido possibilitou a criação de dois grupos de tratamento homogéneos relativamente às distribuições das variáveis, e assim foi possível calcular o "Average Treatment Effect for the Treated (ATT)". Também se provou que os PS ajustados a partir de GAMs originam melhores emparelhamentos, em termos gerais, como sugere a literatura (Woo et al., 2008).

Outra metodologia baseada em Propensity Scores usada foi a Estratificação. Aqui, tomando partido dos Propensity Scores ajustados a partir de GAMs com todas as variáveis registadas como dependentes, agrupou-se os indivíduos em 5 estratos. Dentro de cada estrato, os indíviduos são semelhantes relativamente ao PS e, consequentemente, semelhantes em relação à distribuição das variáveis, pelo que assim é possível calcular o "Average Treatment Effect (ATE)" (Rosenbaum and Rubin, 1984). Outro modo usado para calcular o ATE, foi pela metodologia de "Inverse Probability of Treatment Assignment (IPTW)". O IPTW consiste em atribuir um peso a cada indivíduo, sendo cada peso calculado com base no PS de cada indivíduo. A soma dos pesos em cada grupo de tratamento é igual. Este procedimento cria duas amostras sintéticas com a mesma dimensão, que são diretamente comparáveis, não levantando quaisquer problemas relativos a um possível confundimento (Austin, 2011a).

Ambos os métodos de regressão e as metodologias usando PS, retornaram as mesmas conclusões. O uso do acesso transradial ou transfemoral não influencia a ocorrência de complicações neurológicas peri-procedimento. Assim, o uso clinicamente preferencial do acesso transradial pode continuar a ser aplicado sem preocupações de maior. Estas conclusões reforçam a literatura (Raposo et al., 2015).

Palavras-Chave: Propensity Score, Modelos Aditivos Generalizados, Matching Genético, Estratificação, Inverse Probability of Treatment Weighting

Contents

Abstract i

Resumo v

1 Introduction 1
- 1.1 The Problem . . . 1
  - 1.1.1 Coronary Artery Disease . . . 1
  - 1.1.2 Stroke and Transient Ischemic Attack . . . 1
  - 1.1.3 Diagnostic and Interventional Procedures . . . 2
- 1.2 State of the Art . . . 3
- 1.3 Objectives . . . 4
- 1.4 Work Overview . . . 4

2 Study Design 7
- 2.1 Observational Study . . . 7
- 2.2 Data Source and Entry Criteria . . . 7
- 2.3 Variables . . . 7
  - 2.3.1 Variables Description . . . 7
  - 2.3.2 Sample Baseline Characteristics . . . 9
  - 2.3.3 Definition of Study Outcome . . . 10

3 Methodology 11
- 3.1 Generalized Linear Models . . . 11
  - 3.1.1 Introduction . . . 11
  - 3.1.2 Estimation . . . 12
  - 3.1.3 Hypothesis Tests and Model Selection . . . 13
  - 3.1.4 The Separation Problem and Firth's LR correction . . . 13
  - 3.1.5 Minimum p-value approach . . . 14
  - 3.1.6 Predictive Power . . . 15
  - 3.1.7 Discriminative Power . . . 15
- 3.2 Generalized Additive Models . . . 16
  - 3.2.1 Introduction . . . 16
  - 3.2.2 Additive Models . . . 16
  - 3.2.3 Estimation Algorithm . . . 16
  - 3.2.4 Generalized Additive Models . . . 17
- 3.3 Some Basis in Epidemiology and Propensity Score Methodology . . . 19
  - 3.3.1 The Internal Validity of Observational Studies . . . 19
  - 3.3.2 Confounding . . . 19

3.3.3 Neyman-Rubin's Counterfactual Framework . . . 19
3.3.4 Types of Treatment Effects . . . 21
3.4 The Propensity Score . . . 21
3.5 Propensity Score Matching . . . 22
3.5.1 Traditional Matching . . . 22
3.5.2 Matching on the Propensity Score . . . 22
3.5.3 Variable Selection . . . 23
3.5.4 Balancing Diagnostics . . . 25
3.5.5 Genetic Matching . . . 25
3.5.6 Estimation of the Average Treatment Effect for the Treated . . . 27
3.6 Stratification . . . 28
3.6.1 Estimation of the Average Treatment Effect . . . 28
3.7 Inverse Probability of Treatment Weighting . . . 29
3.7.1 Estimation of the Average Treatment Effect . . . 29
3.8 GLM and PS mixed approach: Covariate Adjustment . . . 30
3.9 Comparison of all Propensity Score Methods . . . 30
3.10 Comparison between Propensity Score and Regression Methods . . . 32

4 Results **33**
4.1 Logistic Regression . . . 33
4.1.1 Univariate Analysis . . . 33
4.1.2 Multivariable Analysis . . . 36
4.1.3 Firth's Correction . . . 38
4.1.4 Treatment Effect Estimation (Odds Ratio) . . . 39
4.2 Matching on the Propensity Score . . . 39
4.2.1 Propensity Score Model (Logistic Regression) and Matching . . . 41
4.2.2 Propensity Score Model and Matching: GAM Regression . . . 44
4.2.3 Genetic Matching . . . 47
4.2.4 Treatment Effect Estimation (Average Treatment Effect for the Treated) . . . 53
4.3 Stratification . . . 53
4.3.1 Treatment Effect Estimation (Average Treatment Effect) . . . 56
4.4 Inverse Probability Treatment Weighting . . . 57
4.4.1 Treatment Effect Estimation (ATE) . . . 57
4.5 GLM and PS mixed approach: Covariate Adjustment . . . 57

5 Discussion **59**

6 Conclusion **62**

Other Work Accomplished **63**

Bibliography **67**

Remissive Index **67**

Appendices **68**

A Logistic Regression **69**

B Firth's Correction **73**

C Propensity Score Matching **75**

D Stratification **84**

E IPTW **88**

F Functions **89**

List of Tables

2.1 Name, Support and Description for each recorded variable. 8
2.2 Sample Baseline Characteristics. Percentage (%) for categorical variables and Average ± Standard Deviation (SD) for continuous variables. 9

3.1 Link Common distributions with typical uses 11
3.2 Two examples of Complete Separation (left) and Quasi-Separation (right), for a sample size=20000 with 25 events. X is the predictor and Y the outcome. 14
3.3 Contingency table for cutpoint c 15
3.4 Advantages and Disadvantages of all PS methods and their respective Treatment Effect Estimated. Table retrieved, partly, from Deb et al. (2015) 31

4.1 Contingency tables for every categorical variable in the dataset regarding each study endpoint group. Average and Standard Deviation(SD) for continuous variables with normal distribution and Minimum(min.)/Median/Maximum(max.) for non-normal continuous variables, regarding each study endpoint group. p-value obtained from the univariate logistic regression (LR). 33
4.2 Odds ratio estimates and the corresponding 95% confidence intervals (CI) from the Univariate logistic regression, regarding the variables under study with a p-value less than 0.25 and the exposure (Transradial Approach). 34
4.3 Minimum p-value approach regarding Quantitative Continuous Variables 36
4.4 Final Model. Estimated coefficients, Odds Ratios, the corresponding 95% Confidence Intervals, the p-values from the Wald Tests, the Hosmer-Lemeshow p-value and the AUC value. 37
4.5 Hosmer-Lemeshow predicted risk bins. For each bin: Mean predicted and mean observed events, total predicted and total observed events. (model from Table 4.4). 38
4.6 Logistic Regression and Firth's Correction univariate analysis regarding β estimates, their standard errors (s.e.) and their p-values. 38
4.7 Firth's Corrected Model. β coefficients, Odds Ratios, the corresponding 95% Confidence Intervals, p-values from the Likelihood Ratio test, p-value from the Hosmer-Lemeshow test and AUC. 39
4.8 Standard Mean Difference (SMD) between Transradial and Transfemoral Group, before matching. Regarding four groups of variables: True Confounders, Potential Confounders, Variables Associated with Treatment Assignment and Other Variables. 40
4.9 Variable Subsets under study. 41

4.10 Standard Mean Difference (SMD) between Transradial and Transfemoral Group, before matching and after matching. Propensity Scores used for matching were estimated by different propensity score models (True Confounders, Potential Confounders, Variables associated with treatment assginment and all observed variables). Matching with Replacement, recurring to NNM-CD with a 0.2 caliper. 42
4.11 Standard Mean Difference (SMD) between Transradial and Transfemoral Group, after matching. Propensity Scores used for matching estimated by different propensity score models (Variables associated with treatment assginment and All observed variables), with and without GAM regression. Matching with Replacement, recurring to NNM-CD with a 0.2 caliper. 46
4.12 Standard Mean Difference (SMD) between Transradial and Transfemoral Group, after Genetic matching (pop. size=300). Propensity Scores estimated by different propensity score models with Variables associated with treatment assginment and All observed variables, modeled by LR and GAM. Matching with Replacement, recurring to NNM-CD with a 0.2 caliper. Loss Function: Maximum Kolmogorov-Smirnov statistic. 50
4.13 Weights regarding Generalized Mahalanobis Distance (GMD), for every Genetic Matching performed [Variable Subset + PS estimates]. SMD for every variable before matching. . 52
4.14 Average Treatment Effect for the Treated and the respective Standard-Deviation (SD) estimated by Matching. 53
4.15 Standard Mean Difference (SMD) between Transradial and Transfemoral Group, in each stratum. Propensity Scores used estimated by a GAM with All observed variables as dependent covariates. 55
4.16 Average Treatment Effect estimates (Stratum-specific and Overall) and its Standard-Deviation (SD), estimated by Stratification 57
4.17 Average Treatment Effect and its Standard-Deviation (SD) estimated by IPTW 57
4.18 LR Model with PS estimates as variable. β coefficients, Odds Ratios, its 95% confidence intervals, and the Wald Test p-values. 58
4.19 Hosmer-Lemeshow predicted risk bins. For each bin: Mean predicted and mean observed events, total predicted and total observed events. Regarding model in Table 4.18. 58

List of Figures

1.1 Example of a narrowed artery due to atherosclerosis. Taken from: Blausen Medical Communications, Inc. 1
1.2 Angiography through the femoral artery depicted. Taken from: Blausen Medical Communications, Inc. 2
1.3 Angioplasty depicted. A: Catheter with the balloon collapsed. B: Catheter with the balloon inflated. C: Widened artery. 3
1.4 (a) Radial Artery depicted (right arm). (b) Femoral Artery depicted (right leg). (c) Schematic of the coronary arteries (around the heart). Taken from: *Anatomy of the Human Body* (Gray, 2000) 4

3.1 Example: Estimated propensity scores $\hat{q}(x)$ distributions in treated and non-treated groups. Represented by a box-plot. 24
3.2 Search for the best Propensity Score model. Figure taken from Diamond and Sekhon (2012). 24
3.3 Representation of Genetic Matching. Figure Taken from Diamond and Sekhon (2012) . . 27

4.1 Univariate GAM functional forms regarding Age, BMI, Systolic BP, Diastolic BP, FLuoroscopy Time and Contrast Volume as covariates and Stroke/TIA as the dependent one. The functional form is represented by the filled lines and the confidence intervals are represented by the dashed lines. 35
4.2 Minimum p-value approach to the quantitative continuous variables. Dashed lines represent the minimum p-value (horizontal), and the corresponding cutpoint (vertical). . 36
4.3 Calibration Plot and ROC curve regarding the model in Table 4.4. 37
4.4 QQ-plots before and after matching regarding Age, Systolic Blood Pressure, Diastolic Blood Pressure and Fluoroscopy Time. (a) QQ-plots before matching. After matching, the LR PS model used to match is (b) the one containing all variables, (c) the one containing all the variables associated with treatment assignment. 43
4.5 Univariate GAM functional forms regarding Age, BMI, Systolic BP, Diastolic BP, FLuoroscopy Time and Contrast Volume as covariates and Transradial Approach as the dependent one. The functional form is represented by the filled lines and the confidence intervals are represented by dashed lines. 44
4.6 ROC plots and Calibration Plots regarding different PS models, with All Variables, All Variables (GAM), Variables Associated with Treatment Assignment, Variables Associated with Treatment Assignment (GAM). AUC and its (CI) are shown within each ROC plot. 45
4.7 QQ-plots after matching regarding Age, Systolic Blood Pressure, Diastolic Blood Pressure and Fluoroscopy Time. The PS estimates used for matching are from: (a) LR containing all variables (b) GAM containing all variables, (c) LR containing the variables associated with treatment assignment, (d) GAM containing the variables associated with treatment assignment. 47

4.8 Histograms regarding Age, BMI, Systolic Blood Pressure, Diastolic Blood Pressure, Fluoroscopy Time, Contrast Volume and Number of Catheters Used. Bar Diagrams regarding Coronary Artery Decease and Number of Treated Segments, ... 48
4.9 Histograms of the: (a) Crude PS estimates and the (b) Transformed PS estimates. Estimated from the LR PS model with All Variables. 49
4.10 QQ-plots after matching regarding Age, Systolic Blood Pressure, Diastolic Blood Pressure and Fluoroscopy Time. The matching method is: (a) genetic containing all variables plus the LR PS estimates, (b) genetic containing all variables plus the GAM PS estimates, (c) genetic containing variables associated with treatment assignment plus the LR PS estimates, (d) genetic containing variables associated with treatment assignment plus the GAM PS estimates. 51
4.11 (a) PS estimates by Stratum, and by treatment group (Transradial coloured Blue and Transfemoral coloured Green). Stratum delimiters coloured Red. (b) Vioplots of the PS estimates of the two treatment groups. Vioplots are box-plot analogues which depict, additionally, the density function (coloured Yellow). 54
4.12 QQ-plots after stratification regarding Age, Systolic Blood Pressure, Diastolic Blood Pressure and Fluoroscopy Time, at each stratum. Stratum 1: Subjects within 0-1^{st} quintile, Stratum 2: Subjects within 1^{st}-2^{nd} quintile, Stratum 3: Subjects within 2^{nd}-3^{rd} quintile, Stratum 4: Subjects within 3^{rd}-4^{th} quintile, Stratum 5: Subjects within 4^{th}-5^{th} quintile. 56
4.13 (a) ROC curve regarding model in Table. b) Calibration Plot regarding the Hosmer-Lemeshow Test applied to model in Table 4.18 58

6.1 Screenshot of the "Propensity Scores: An application to Medicine" application. 63

List of Abbreviations

AM Additive Model

ATE Average Treatment Effect

ATT Average Treatment Effect for the Treated

BP Blood Pressure

CABG Coronary Artery Bypass Graft

CAD Coronary Artery Disease

GAM Generalized Additive Model

GLM Generalized Linear Models

GM Genetic Matching

GMD Generalized Mahalanobis Distance

IPTW Inverse Probability of Treatment Weighting

ITAA Ignorable Treatment Assignment Assumption

LR Logistic Regression

MD Mahalanobis Distance

MLE Maximum-Likelihood Estimation

NNM-CD Nearest Neighbour Matching with a Specied Caliper Distance

NRCF Neyman's Rubin Counterfactual Framework

OR Odds Ratio

PAD Peripheral Artery Disease

PCI Percutaneous Coronary Intervention

PS Propensity Score

RCT Randomized Controlled Trial

ROC Receiver Operating Charateristic

SUTVA Stable Unit Treatment Value Assumption

TIA Transient Ischemic Attack

Chapter 1

Introduction

1.1 The Problem

In cardiology, the radial artery has recently been used with increasing frequency as the preferential vascular access both in diagnostic and interventional procedures. This so called transradial approach has risen some concerns relative to the higher risk of peri-procedural neurologic complications like Stroke and TIA (Raposo et al., 2015). In this work, the association between this approach and possible further neurologic complications is studied.

1.1.1 Coronary Artery Disease

One of the main reasons persons are subjected to invasive techniques, involving the manipulation of the ascending aorta and subclavian arteries, is the Coronary Artery Disease. Also known as ischemic heart disease, it is characterized by the shortage of blood flow to the heart provoking damage to it (e.g. Myocardial Infarction(MI)). The underlying mechanism responsible for this involves atherosclerosis of the arteries of the heart. Atherosclerosis is the process responsible for the thickening of artery walls due to accumulation of white blood cells and remnants of dead cells including cholesterol and triglycerides. This provokes narrowing and loss of elasticity of the artery walls, impairing the blood flow (Figure 1.1)(National Heart and Institute, 2016). The main risk factors are high blood pressure, dyslipidemia (abnormal amount of lipids in the blood), smoking, diabetes, lack of exercise, poor diet and alcohol consumption. Peripheral artery disease (PAD) is an analogue disease, concerning only the narrowing of arteries that are not coronary (Mendis; Shanthi et al., 2011).

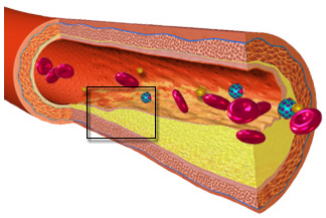


Figure 1.1: Example of a narrowed artery due to atherosclerosis. Taken from: Blausen Medical Communications, Inc.

1.1.2 Stroke and Transient Ischemic Attack

Stroke, also known as cerebrovascular accident (CVA) is a consequence of poor blood flow to the brain resulting in cell death, affecting the regular brain functioning. When a stroke occurs, the most common

signs are inability to move or feel parts of the body, difficulties in speaking and understanding and partial loss of vision. If these symptoms last only for a few hours it is known as a Transient Ischemic Attack (TIA), otherwise it is a Stroke. The consequences of Stroke or TIA can be permanent. The main risk factors are high blood pressure, smoking, obesity, high blood cholestrol and diabetes mellitus . Typically, the two main causes of CVA are the blockage of an artery and cerebral bleeding (Donnan et al., 2008).

In the context of this work, the study endpoint will be the Stroke or TIA provoked by the manipulation unavoidably associated with invasive procedures.

1.1.3 Diagnostic and Interventional Procedures

As stated before, the cardiac procedures, may have diagnostic or interventional purposes. The need for catheterization is common to both situations. A catheter is an device which is guided through the vascular access point to the coronary arteries and can be used to perform a wide range of procedures. The entrance site of the catheter is usually the radial (Transradial Approach) or the femoral artery (Transfemoral Approach) (Figure 1.4). An overview through the most common invasive interventional and diagnostic procedures is presented.

1.1.3.1 Diagnostic Procedures

Angiography The most common diagnostic procedure. Angiography is a medical imaging technique used to get images from the heart chambers, arteries or veins. This is performed by injecting a dose of contrast volume and then, by applying X-ray techniques like Fluoroscopy, obtaining a view of a given anatomical structure's morphology, trough opacification of its lumen. A depiction of angiography is displayed in Figure 1.2 (Topol and Teirstein, 2016).

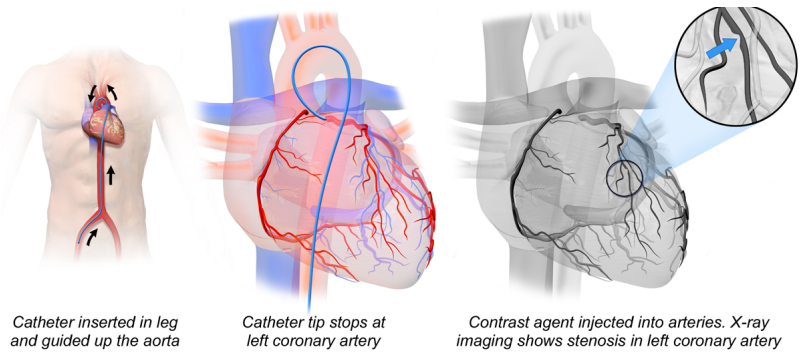


Figure 1.2: Angiography through the femoral artery depicted. Taken from: Blausen Medical Communications, Inc.

Catheters can also be used to guide other diagnostic devices, like for example intravascular ultrasound probes, pressure and/or flow sensor-tipped guide wires.

1.1.3.2 Interventional Procedures

Many interventional procedures are performed percutaneously. **Angioplasty**, also known as Percutaneous Coronary Intervention (PCI), is by far the most common. In this procedure the main objective is to widen the obstructed arteries (or veins). Here, the catheter has a collapsed balloon at its tip, and when it reaches the obstructed localtion, the balloon is inflated by the operator and pushes back the artery walls and its plaques widening them (Figure 1.3). This improves the blood flow as pretended. Balloons can also be

used to dilate stenotic valves (**balloon valvuloplasty**) and to implant stents (Topol and Teirstein, 2016).

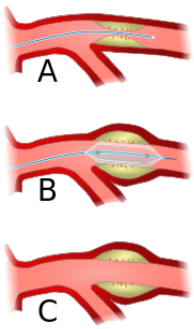


Figure 1.3: Angioplasty depicted. A: Catheter with the balloon collapsed. B: Catheter with the balloon inflated. C: Widened artery.

A stent is a metal or polymeric ”tube” used to keep an artery widened. Stenting is an upgrade to plain-balloon angioplasty and increases both its safety and efficacy in many clinically important situations. Coronary angioplasty with stenting is widely used and over the years it has replaced Coronary Artery Bypass Grafting (CABG) - a surgical procedure - in many patients warranting myocardial revasculzarization (Topol and Teirstein, 2016).

1.2 State of the Art

Many forms of cardiac and vascular disease that may need invasive techniques to be managed, particularly, obstructive coronary artery disease. These invasive techniques can have therapeutic or merely diagnostic purposes. In both situations, there is the need of an arterial access to perform the procedure. During the second half of the 20^{th} century, various forms of arterial access were developed like the access through the radial, ulnar, sub-clavian and femoral arteries, performed using the percutaneous Seldinger technique or direct surgical exposure (Bourassa, 2005). Until recent years, the Transfemoral Approach (procedure that uses the femoral artery as access point) was the preferred one by interventional cardiologists due to the high compatibility with the Seldinger puncture technique and due to the large calibre of the vessel allowing various sheaths. Although Transfemoral Approach gained great popularity, it has been proved that this approach causes many peri-procedural complications like bleeding, and therefore is responsible for longer admission times and higher morbidity and mortality (Nathan and Rao, 2012; Burzotta et al., 2013).

The anatomic characteristics of the radial artery make it a great access point for many procedures (not only cardiac), e.g. it is accessible in the anterior aspect of the forearm, the vessel calibre is large enough and vascular complications are rare (Bhat et al., 2012) (Figure 1.4). Many studies have pointed that Transradial Approach has many benefits in diverse clinical settings like reduced mortality, particularly, in the group of patients with acute coronary syndromes, in whom bleeding can nullify the benefits of the procedure itself (Nathan and Rao, 2012). In addtion, Transradial Approach is less expensive (early and mid-term) (Mitchell et al., 2012). Improvements in access materials and catheters dedicated to this technique have contributed to the wide spread of radial artery as access point. This choice do not impact the overall procedure performance (Dangoisse et al., 2013).

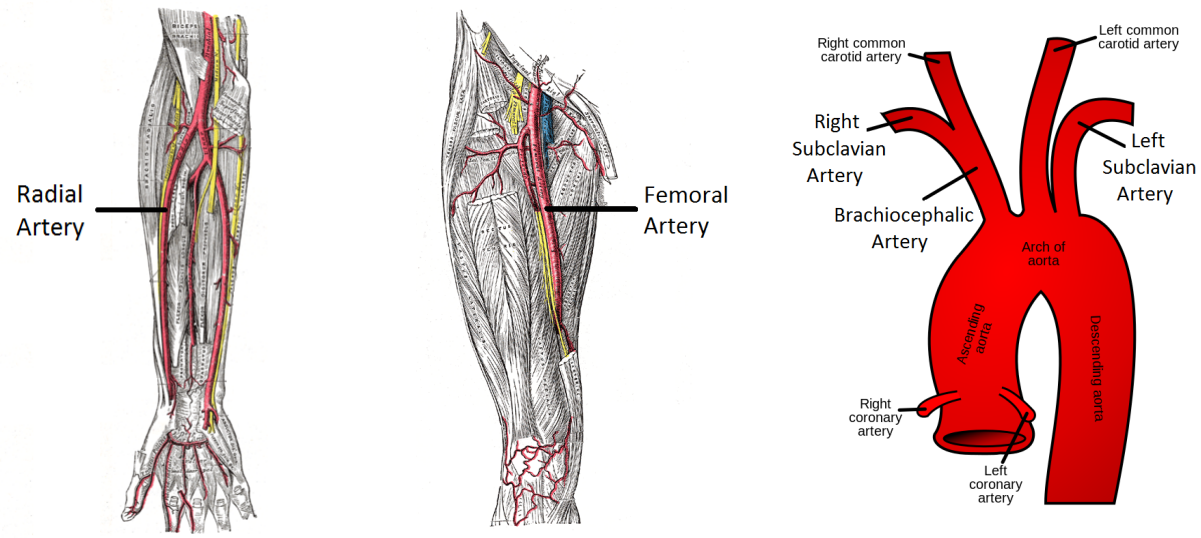


Figure 1.4: (a) Radial Artery depicted (right arm). (b) Femoral Artery depicted (right leg). (c) Schematic of the coronary arteries (around the heart). Taken from: *Anatomy of the Human Body* (Gray, 2000)

These access points (radial and femoral) can impact the way ascending aorta and subclavian arteries (specially the right brachiocephalic artery) are manipulated (Figure 1.4). Because of this, there are concerns regarding the potentially increased risk of cerebral embolization. Recently, a significantly higher rate of micro-embolization within a subset of patients undergoing Transradial Approach has been reported (Jurga et al., 2011). However, these findings have not been confirmed in a large multicenter registry (Ratib et al., 2013b). Due to the low number of peri-procedural Strokes/Transient Ischemic Attacks (TIA), it is very difficult to design and perform powered studies, in order to identify significant differences regarding this important endpoint.

In 2015, Raposo *et al* reported that there were no increased risk of peri-procedural Stroke/TIA by undergoing catheterization through the radial artery. This work uses this very same data and tries to extend the previous work by applying other type of PS methodologies.

1.3 Objectives

The objective of this study is to study the association between the vascular access used (arterial or femoral artery) in a cardiac procedure and the occurrence of peri-procedural neurological complications (Stroke or Transient Ischemic Attack). This association is to be studied by regression methods (logistic regression) and Propensity Score Analysis (Matching, Genetic Matching, Stratification and IPTW).

1.4 Work Overview

Chapter 1 is a brief introduction to the problem. This includes several medical explanations about the problem, the state-of-the-art, the objectives and why this work is important. In Chapter 2, it is explained how the study was designed, the variable description, the sample baseline characteristics and the definition of the study outcome. In Chapter 3, all the statistical methodologies employed are described in detail, namely, Generalized Linear Models, Generalized Additive Models, the Propensity Scores properties

and its applications like Matching, Stratification and IPTW. The results of the application of these methodologies to the data are reported in Chapter 4. The results are discussed in Chapter 5, and Chapter 6 contains the final Conclusions regarding the problem risen in Chapter 1.

Chapter 2

Study Design

2.1 Observational Study

The main objective of observational studies is to establish causality between the hypothetical cause (in clinical research, commonly described as ”treatment” or ”exposure”), and the effect. Usually, there are two groups in study, the control and the treatment group, which differ only in the treatment assigned. The main objective of a study is to prove that the outcome difference between groups is caused by the type of treatment received (Gordis, 2014). In an observational study, the treatment assignment is not randomized, consisting on extracting knowledge from databases already established.

2.2 Data Source and Entry Criteria

This observational study was performed using data from the Angiography and Cardiovascular InteRventiOn centre at Santa Cruz HoSpital (ACROSS) Registry. The database (Cardiobase ®, Infortucano ®, Lisbon, Portugal) has records regarding anthropometric, demographic, clinical and procedure related variables. Initially, all subjects who were exposed to invasive procedures involving an arterial vascular access, during January'2006 until November'2012, were screened (n=19,961). The entry criteria were the following:

- age higher or equal than 18 years old;
- being subjected to interventional or diagnostic procedures involving any manipulation of the ascending aorta or the aortic arch.

The final study population size is 16,710 patients.

Radial Artery was used as primary vascular access site in 4,195 cases (25% of the sample) and in 36 additional patients after conversion from femoral-to-radial. Thus, the final transradial group size is 4,231.

2.3 Variables

2.3.1 Variables Description

Variables Description for each recorded variable is displayed in Table 2.1:

Table 2.1: Name, Support and Description for each recorded variable.

Variable	Support	Description
Demographic		
Age (years)	{18,19, 20,...}	Age in years
Male Gender	{0,1}	1:Male 0:Female
Body Mass Index (BMI) (Kg/m^2)	(0,+∞)	The body mass index value
Cardiovascular Risk Factors		
Diabetes mellitus	{0,1}	1:Has Diabetes mellitus 0:Do not have Diabetes mellitus
Hypertension	{0,1}	1:Has Hipertension 0:Do not have Hypertension
Smoking (current or former)	{0,1}	1:Was/Is a smoker 0:No smoking history
Dyslipidemia	{0,1}	1:Has Dyslipidemia 0:Do not have Dyslipidemia
Prior Clinical History		
Myocardial Infarction (MI)	{0,1}	1:MI history 0:No MI history
Percutaneous Coronary Intervention (PCI)	{0,1}	1:PCI history 0:No PCI history
Coronary Artery Bypass Surgery (CABG)	{0,1}	1:CABG history 0:No CABG history
Peripheral Artery Disease (PAD)	{0,1}	1:PAD history 0:No PAD history
Stroke/Transient Ischemic Attack (TIA)	{0,1}	1:Stroke/TIA history 0:No Stroke/TIA history
Non-CABG Surgery	{0,1}	1: Non-CABG Surgery History 0: No Non-CABG Surgery History
Moderate/severe Chronic Renal Disease (CRD)	{0,1}	1:CRD history 0:No CRD history
Renal Transplant	{0,1}	1:Renal Transplant history 0:No Renal Transplant history
Clinical Setting and Procedural Characteristics		
Interventional Procedure	{0,1}	1:Angioplasty/Percutaneous coronary intervention 0:Cardiac catheterization for diagnostic purposes
Acute Coronary Syndrome	{0,1}	1:Has Acute Coronary Syndrome 0:Do not have Acute Coronary Syndrome
Aortic Valvulopathy	{0,1}	1: Has Aortic Valvulopathy 0: Do not have Aortic Valvulopathy
Systolic Blood Pressure (SBP) (mmHg)	(0,+∞)	The SBP value in mercury milimeters (mmHg)
Diastolic Blood Pressure (DBP) (mmHg)	(0,+∞)	The DBP value in mercury milimeters (mmHg)
Coronary artery disease extension	{0,1,2,...}	Number of vessels with stenosis*
Number of treated segments	{0,1,2,...}	Number of vessels treated. Only refers to patients intervened with angioplasty
Fluoroscopy time (min)	(0,+∞)	The Fluoroscopy** duration in minutes
Contrast Volume (mL)	{0,1,2,...}	The quantity of contrast used during fluoroscopy in milliliters

Number of Catheters Used	{0,1,2,...}	The number of Catheters used during the procedure

* abnormal narrowing in a blood vessel or other tubular organ or structure.

** imaging technique used during procedure.

2.3.2 Sample Baseline Characteristics

The sample in study is, on average, 65.79 years old, is predominantly male (67%), 28% is diabetic, 73% is hypertensive, 40% is/was a smoker and 60% has dyslipidemia. More Sample Baseline characteristics are summarized in Table 2.2.

Table 2.2: Sample Baseline Characteristics. Percentage (%) for categorical variables and Average ± Standard Deviation (SD) for continuous variables.

Variable	
Demographic	
Age (years) [average ± SD]	65.79 ± 11.59
Male Gender (*n*, %)	11 134 (67%)
BMI (Kg/m^2) [average± SD]	27.35 ± 4.04
Cardiovascular Risk Factors	
Diabetes mellitus (*n*, %)	4 703 (28%)
Hypertension (*n*, %)	12 211 (73%)
Smoking (current or former) (*n*, %)	6 601 (40%)
Dyslipidemia (*n*, %)	10 039 (60%)
Prior Clinical History	
MI (*n*, %)	3 569 (21%)
PCI (*n*, %)	3 930 (24%)
CABG (*n*, %)	1 537 (9%)
Stroke/TIA (*n*, %)	10 284 (8%)
PAD (*n*, %)	1 112 (7%)
Non-CABG Surgery (*n*, %)	425 (3%)
Moderate/severe CRD (*n*, %)	531 (3%)
Renal Transplant (*n*, %)	70 (0,4%)
Clinical Setting and Procedural Characteristics	
Interventional Procedure (*n*, %)	6 973 (42%)
Acute Coronary Syndrome (*n*, %)	6 204 (37%)
Aortic Valvulophaty (*n*, %)	1 220 (7%)
SBP (mmHg) [average± SD]	139.15 ± 28.16
DBP (mmHg) [average± SD]	72.27 ± 13.05
Coronary artery disease extension [average± SD]	1.35 ± 1.14
Number of Treated Segments [average± SD]	1.35 ± 1.15
Fluorosopy time (min) [average± SD]	9.21 ± 0.25
Contrast Volume (ml) [average± SD]	136.65 ± 95.60
Number of Catheters Used [average± SD]	2.80 ± 1.36

2.3.3 Definition of Study Outcome

The study outcome was defined as the occurrence of clinically significant stroke or transient ischemic attack (TIA) related to endovascular manipulation. At the hospital, all cases suspected of neurological complications are submitted to brain imaging and recorded in the hospital database. In order to identify possible Stroke/TIA related to the procedure, it was created a shortlist of patients who underwent both endovascular manipulation and brain imaging. Then, each one of the cases on the shortlist were reviewed, preferably by a neurologist, to access if the Stroke/TIA was due to the procedure. These confirmed cases had to occur within 48 hours of the procedure, and had to be unrelated to coronary artery bypass grafting. For patients discharged or transferred to other hospitals within this 48 hour period, their data was registered in the database: Cardiobase ®, Infortucano, Lisbon, Portugal.

Chapter 3

Methodology

3.1 Generalized Linear Models

3.1.1 Introduction

A regression analysis is a way to access the relationship between different independent variables $(X_1, X_2, ..., X_k)$ and a variable of interest Y. A generalized linear model (GLM) is an extension of the linear classic model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$, $\epsilon \cap N(0, \sigma^2)$, where each β_j $(j = 1, ..., k)$ express the contribution of each variable to the value of Y (Amaral Turkman and Loiola Silva, 2000).

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{3.1}$$

A GLM is a regression model where Y belongs to the exponential family, and a linear structure $(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$ is related to $E(Y|\mathbf{X}) = \mu$ by a link function $g(\cdot)$ as in Equation 3.1. The link function used depends, mostly, on the Y response type. In Table 3.1 are described some link functions:

Table 3.1: Link Common distributions with typical uses

Distribution of Y	**Support of distribution**	**Link Name**	**Link Function**$(g(\mu))$
Normal	real$(-\infty, +\infty)$	identity	μ
Exponential or Gamma	real$(0, +\infty)$	inverse	$-\mu^{-1}$
Poisson	integer $(0, 1, 2, \ldots, +\infty)$	log	$ln(\mu)$
Bernoulli	integer $\{0, 1\}$	logit	$ln(\frac{\mu}{1-\mu})$

Suppose Y is a binary dependent variable and $\mathbf{X} = (X_1, X_2, X_3, ..., X_k)$ the independent variables that may be asssociated with Y. So, for each subject *i*, these are the observations: $(y_i, x_{i1}, x_{i2}, x_{i3}, ..., x_{ik})$. Formalizing, $Y_i \sim Bernoulli(p_i)$, $\mu_i = E(Y_i|\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})) = p_i = Pr(Y_i = 1|\mathbf{x}_i)$, $i = 1, 2...n$.

Thus, the appropriate link function is the logit because $\hat{\mu}$ will always fall within 0 and 1. The GLM regression that uses the logit link, it is called **Logistic Regression** (LR) and it is represented this way:

$$\begin{aligned} p_i &= \frac{exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}{1 + exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})} \\ ln(\frac{p_i}{1 - p_i}) &= \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} \end{aligned} \tag{3.2}$$

3.1.2 Estimation

The main objective of LR is to estimate the contribution of each variable X for the final outcome and, therefore, every p_i. By the maximum-likelihood estimation method, it is possible to estimate the β coefficients.

The **Maximum-Likelihood Estimation (MLE) method** finds the $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)$ values, which maximize the log-likelihood function:

$$\begin{aligned} l = lnL(\mathbf{y}, \boldsymbol{\beta}) = ln \prod_{i=1}^{n} P(Y_i = y_i | \mathbf{X}_i = \mathbf{x}_i) = ln \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i} = \sum_{i=1}^{n} ln(p_i^{y_i} (1-p_i)^{1-y_i}) \\ = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}) - ln(1 + (\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}))] \end{aligned} \tag{3.3}$$

In order to maximize it:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = 0 \Leftrightarrow U(\boldsymbol{\beta}) = 0 \Leftrightarrow \begin{bmatrix} U(\beta_0) \\ U(\beta_1) \\ \vdots \\ U(\beta_k) \end{bmatrix} = 0 \tag{3.4}$$

$$U(\beta_j) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial p_i} \frac{p_i}{\beta_j} = \sum_{i=1}^{n} \frac{y_i - p_i}{p_i(1-p_i)} x_{ji} p_i (1-p_i) = \sum_{i=1}^{n} x_{ji}(y_i - p_i) \tag{3.5}$$

The $\boldsymbol{\beta}$ parameters that maximize the log-likelihood function, are the ones where $U(\boldsymbol{\beta}) = 0$ (3.4). Since there is no analytical solution for this equation, an interative method like Newton-Raphson (NR) is required.

Let $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ 1 & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{bmatrix}$, $\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix}$,

so $U(\boldsymbol{\beta}) = \mathbf{U} = \mathbf{X}^T(\mathbf{Y\text{-}p})$. In order to find the β parameters where $U(\boldsymbol{\beta}) = 0$, the NR iterative method is applied till convergence is obtained:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{m+1} = \hat{\boldsymbol{\beta}}_m + \left[-\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta}} \right]^{-1}_{(\hat{\boldsymbol{\beta}}_m)} \mathbf{U}_{(\hat{\beta}_m)} \\ \hat{\boldsymbol{\beta}}_{m+1} = \hat{\boldsymbol{\beta}}_m + (\boldsymbol{X^T W X})^{-1}_{(\hat{\boldsymbol{\beta}}_m)} \boldsymbol{X^T}(\mathbf{Y\text{-}p})_{(\hat{\boldsymbol{\beta}}_m)} \end{aligned} , \tag{3.6}$$

where $\mathbf{W} = Diag(\mathbf{p}(1-\mathbf{p}))$

Odds Ratio(OR) is a conditional effect and it is one of the most used risk measures in clinical research. The OR is the ratio between the odds of an "exposed" subject i ($\frac{p_i}{1-p_i}$) and the odds of a "non-exposed" subject j ($\frac{p_j}{1-p_j}$), holding all the other variables constant. Suppose, for two subjects i and j, $x_{i2} = x_{j2}$, $x_{i3} = x_{j3}$, ..., $x_{ik} = x_{jk}$ and $x_{i1} - x_{j2} = 1$. In this case, the subject i is the one being "exposed" to the risk factor X_1. So, assuming a logistic model, the odds ratio is given by e^{β_1}. This is proved here:

$$
\begin{aligned}
log(\frac{p_i}{1-p_i}) - log(\frac{p_j}{1-p_j}) &= (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) - (\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk}) \\
log(\frac{\frac{p_i}{1-p_i}}{\frac{p_j}{1-p_j}}) &= (\beta_0 - \beta_0) + (\beta_1 x_{i1} - \beta_1 x_{j1}) + (\beta_2 x_{i2} - \beta_2 x_j 2) + \cdots + (\beta_k x_{ik} - \beta_k x_{jk}) \\
log(OR) &= \beta_1 (x_{i1} - x_{j1}) \\
OR &= e^{\beta_1}
\end{aligned}
\tag{3.7}
$$

3.1.3 Hypothesis Tests and Model Selection

Wald Test: This test is used to test, for example, $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ (*j*=1,...,k). The covariance matrix of $\hat{\boldsymbol{\beta}}$ is : $Cov(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X^T W X})^{-1}$, so the standard deviation of $\hat{\beta}_j$ is $sd(\hat{\beta}_j) = \sqrt{(\boldsymbol{X^T W X})^{-1}_{(j+1,j+1)}}$. As $\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1)$, it is possible to make inferences about the $\boldsymbol{\beta}$ values.

Likelihood-ratio test: In order to compare the goodness-of-fit of 2 models, the Likelihood-ratio test is performed. This test evaluates if the data is more likely under one model or another.

Assume a logistic model adjusted for k variables (M_k), and another one adjusted for the same k variables plus 1 (M_{k+1}). In this case, the likelihood ratio is $\lambda = \frac{L(M_k)}{L(M_{k+1})}$. The model with more variables (M_{k+1}) will always fit at least as well, i.e., have a greater or equal likelihood, than the model with less variables (M_k), so $log(\lambda) < 0$. M_{k+1} has n-(k+1)-1 free parameters, while M_k has n-k-1 free parameters

Formalizing the hypothesis, $H_0 : \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ vs $H_1 : \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k, \beta_{k+1})$. As $-2log(\lambda) \sim \chi^2_{(n-k-1)-(n-(k+1)-1)}$, the significance is calculated this way. If there is no statistical difference between the models (H_0 is not rejected), the model with less parameters is the chosen one. Other statistical tests can be used like the **Score test**.

Aikaike's Information Criterion (AIC): AIC measures the relative quality of a model in comparison with other models, given the same dataset. This measure estimates the information lost by the model representation of the data. So a better model will have a lower AIC. The big question here is: In order to get a better adjustment, Is the complexity of the model worth? If two models have the same likelihood, given the same dataset, the one with less AIC will be the one with less variables. Inversely, between two models with the same number of variables, the one with less AIC is the one with a higher likelihood. Basically, this measure deals with the compromise between few variables and a great likelihood. AIC is defined in Equation 3.8:

$$
AIC = 2k - 2ln(L) \tag{3.8}
$$

where k is the number of variables and L the model's likelihood.

3.1.4 The Separation Problem and Firth's LR correction

The Separation Problem When few events occur in a dataset (Y=1), there is great risk of facing the "separation" or the "monotone likelihood" problem . This kind of problem arises when the predictor (or a linear combination of some subset of the predictors) is associated with only one outcome value (highly predictive). For example, if the predictor X is dichotomous, and the outcome y = 1 for all observed x = 1, and y=0 for all x=0, the condition "complete separation" is said to be obtained. If the outcome

y=1 for all observed x=1, but there are different outcomes values (y=0 and y=1) for x=0, it is called a "quasi-separation" situation (Table 3.2).

Table 3.2: Two examples of Complete Separation (left) and Quasi-Separation (right), for a sample size=20000 with 25 events. X is the predictor and Y the outcome.

		Y	
		0	1
X	0	19975	0
	1	0	25

		Y	
		0	1
X	0	15000	0
	1	4975	25

This separation situations cause problems estimating the coefficients, using the MLE. In a complete separation context, the $\boldsymbol{\beta}$ estimates tend to be infinite because the MLE, continuously, maximizes the likelihood. In a quasi-separation, the β estimates are not infinite but are very large, with huge standard errors. Of course, these estimates are not correct and do not reflect the real effect of X. As the $\boldsymbol{\beta}$ s.e. are large, the Wald's test always returns non-significant statistics. And, the Likelihood Ratio test is not trustworthy as well because the estimated likelihood cannot be taken seriously. To mitigate this problem, a penalized MLE was developed by Firth in 1993 (Firth, 1993).

Firth's Logistic Regression estimation algorithm

In order to reduce this small sample bias, Firth suggested to maximize a penalized maximum likelihood instead:

$$\begin{aligned} L(\boldsymbol{\beta})^* &= L(\boldsymbol{\beta})|I(\boldsymbol{\beta})|^{1/2} \\ logL(\boldsymbol{\beta})^* &= logL(\boldsymbol{\beta}) + 1/2log|I(\boldsymbol{\beta})| \end{aligned} \tag{3.9}$$

being $I(\boldsymbol{\beta}) = E\left[\left(\frac{\partial}{\partial\boldsymbol{\beta}} log(f(X;\boldsymbol{\beta}))\right)^2 |\boldsymbol{\beta}\right]$ the information matrix at β. So the score function is:

$$\begin{aligned} U(\beta_j)^* &= \frac{\partial logL(\boldsymbol{\beta})^*}{\partial\beta_j} \quad (j = 1, \ldots, k) \\ U(\beta_j)^* &\equiv U(\beta_j) + 1/2\, trace[I(\boldsymbol{\beta})^{-1}\{\partial I(\boldsymbol{\beta})/\partial\beta_j\}] \\ U(\beta_j)^* &= \sum_{i=1}^{n}\{y_i - p_i + h_i(1/2 - p_i)\}x_{ir} = 0 \end{aligned} \tag{3.10}$$

being h_i the ith diagonal elements of matrix $\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}$. As $U(\boldsymbol{\beta})$ has no analytical solution, the estimates are obtained iteratively until convergence:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^m + I^{-1}(\boldsymbol{\beta}^{(m)})U(\boldsymbol{\beta}^{(m)})^* \tag{3.11}$$

In this context, the β confidence intervals (CI) are not estimated based on $Cov(\hat{\boldsymbol{\beta}})$, but based on each β profile likelihood. Although the Wald's Test is possible under these circumstances, the Likelihood Ratio test is the test usually used to check if one variable holds within the model. For further information about this method consult Heinze and Schemper's article "A solution to the problem of Separation in Logistic Regression" (Heinze and Schemper, 2002).

3.1.5 Minimum *p*-value approach

In clinical context, often, there is the need to categorize variables in order to improve decision making. For example, if a variable is categorized, it is much easier to make treatment recommendations, determining study eligibility or to balance treatment groups in a clinical trial. (Mazumdar and Glassman, 2000)

Table 3.3: Contingency table for cutpoint c

	Y	
	0	1
X ≤ c	n_{11}	n_{12}
X > c	n_{21}	n_{22}

In order to find the cutpoint that best categorize some variable, the minimum p-value approach was developed. The best cutpoint (if defining two groups) is the one that creates the most different groups regarding their risk of suffering the outcome (Y). This is accessed by finding the cutpoint c that produces the minimum p-value when performing a chi-squared test (Table 3.3). The higher the dependence between the categorized and the outcome variable, the more meaningful risk differences exist between the categorized groups.

A categorized variable can be significant while its respective continuous analog is not. This means that categorization widens the range of possible variables to be included in a model (Mazumdar and Glassman, 2000).

3.1.6 Predictive Power

Hosmer-Lemeshow Test: This comparison is done within subgroups of the model population. The groups are, typically, defined by the deciles of the fitted values of the model. Consider G the number of groups, O_g the number of observed events in group g, E_g the number of expected events in group g, N_g the number of subjects in group g and π_g the predicted risk for the group g. Under H_0 the model is fit. Thus the test statistic is:

$$H = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{N_g \pi_g (1 - \pi_g)} \sim \chi^2_{G-2} \tag{3.12}$$

Calibration Plots: This kind of plots represent, graphically, how close the estimates are from the observed values. Thus, it plots the fitted outcome values against the observed ones. Within LR context, this plotting may be not very pleasant as in a regression with continuous outcome values. Because of this, this kind of representations within LR context, plots the mean of fitted values against the mean of observed values, for each subgroup like in the Hosmer-Lemeshow Test.

3.1.7 Discriminative Power

Receiver Operating Characteristic (ROC): or ROC curve is a graphical feature that analyses the level of discrimination of a binary classifier system (in this case, a LR). This curve measures the ability to differentiate between Y=0 and Y=1. The fitted values $\hat{Y}$ range between 0 and 1, but the real Y values are $\{0,1\}$. There is the need to specify a cut-off value (c), in which one assumes $\hat{Y} = 1$ if $\hat{Y} > c$ and $\hat{Y} = 0$ if $\hat{Y} < c$.

For each c, there is a sensitivity value ($P(\hat{Y} = 1|Y = 1)$ estimated by $\frac{\#\{\hat{y}=1, y=1\}}{\#\{y=1\}}$) and a specificity value ($P(\hat{Y} = 0|Y = 0)$ estimated by $\frac{\#\{\hat{y}=0, y=0\}}{\#\{y=0\}}$). The ROC curve is the representation of (x=1-specificity, y=sensivity) for each value of c. The area under the ROC curve (AUC) is a discrimination measure, which is defined as $AUC = \frac{\#\{correct-discriminations\}}{\#\{total-discriminations\}}$.

3.2 Generalized Additive Models

3.2.1 Introduction

Analyzing the expressions for the GLM (Equation 3.1), one realize that these impose a linear relation between the independent variables and the dependent one. This lack of flexibility is a strong restriction which, often, does not represent the true relationship between variables. This problem can be solved by applying additive and generalized additive models. This type of models express the contribution of each independent variable as a functional form defined through smoothing techniques. The first reference to this kind of models was in 1947 (Leontief, 1947), but they were only brought to the spotlight by the end of the eighties (Buja et al., 1989; Hastie and Tibshirani, 1990).

3.2.2 Additive Models

An additive model is defined through the following expression:

$$Y = \alpha + \sum_{j=1}^{k} f_j(X_j) + \epsilon \tag{3.13}$$

in which:

- The error term ϵ is independent from the covariates $\mathbf{X}=(X_1, \ldots, X_k)$, $j = 1, \ldots, k$.
- $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$
- The partial functions f_j, $j = 1, \ldots, k$ are :
    - arbitrary functions, in most cases univariate functions
    - smooth functions
    - $E[f_j(X_j)] = 0 \Rightarrow E(Y) = \alpha$

3.2.3 Estimation Algorithm

The estimation of each functional form (f_j) regarding each independent variable ($j = 1, \ldots, k$), is done by the *backfitting* algorithm. First introduced in 1981 (Friedman and Stuetzle, 1981), its theoretical basis is represented by Equation 3.14:

$$E[Y - \alpha - \sum_{j \neq k} f_j(X_j) | X_k] = f_k(X_k) \tag{3.14}$$

Suppose there is the need to estimate the model $E(Y|X_1, X_2) = f_1(X_1) + f_2(X_2)$. Let's see how it works from an informal point of view. From the initial estimate $\hat{f}_1(X_1)$, one can estimate $f_2(X_2)$ through the smoothing of the residual $Y - \hat{f}_1(X_1)$ on X_2. Next, a better estimate of $f_1(X_1)$ is obtained smoothing $Y - \hat{f}_2(X_2)$ on X_1. This process will continue till $\hat{f}_1(X_1)$ and $\hat{f}_2(X_2)$ do not suffer major changes in two consecutive iterations.

Formalizing the estimation of the partial functions $\mathbf{f}_j$ $j = 1, \ldots, k$:

$$\begin{aligned}
\mathbf{f}_1 &= \mathbf{S}_1(\mathbf{y} - . - \mathbf{f}_2 - \cdots - \mathbf{f}_k) \\
\mathbf{f}_2 &= \mathbf{S}_2(\mathbf{y} - \mathbf{f}_1 - . - \cdots - \mathbf{f}_k) \\
&\vdots = \vdots \\
\mathbf{f}_k &= \mathbf{S}_k(\mathbf{y} - \mathbf{f}_1 - \mathbf{f}_2 - \cdots - .)
\end{aligned} \tag{3.15}$$

being $\{\mathbf{f}_j = \{f_j(x_{1j}), \ldots, f_j(x_{nj})\}^T$ with $j = 1, \ldots, k$, $\mathbf{S}_j$ corresponds to the smoothing matrix on X_j and the points (.) indicate the missing term in each row. This equation system can be written in the following matrix system:

$$\begin{pmatrix} \mathbf{I} & \mathbf{S}_1 & \ldots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \ldots & \mathbf{S}_2 \\ \vdots & & \ddots & \vdots \\ \mathbf{S}_k & \ldots & \mathbf{S}_k & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_k \end{pmatrix} = \begin{pmatrix} \mathbf{S}_1\mathbf{y} \\ \mathbf{S}_2\mathbf{y} \\ \vdots \\ \mathbf{S}_k\mathbf{y} \end{pmatrix}$$

This matrix system is solved using the *backfitting* algorithm. This way, one will obtain the estimates of $\mathbf{f}_j$, for $j = 1, \ldots, k$.

Considering a sample of *n* subjects, their responses are **y**=$(y_1, \ldots, y_n)^T$, and their observed covariate values **x** are:

$$\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & & \ddots & \vdots \\ x_{n1} & \ldots & x_{n(k-1)} & x_{nk} \end{pmatrix}$$

The *backfitting* algorithm takes the following steps:

1. **Initialization:** Initialize $\hat{\alpha} = \bar{y}$ and $\hat{\mathbf{f}}_j^{(0)} = 0$, or perform a linear regression Y on X_j, to obtain $\hat{\mathbf{f}}_j^{(0)} = \{\hat{f}_j^{(0)}(x_{1j}), \ldots, \hat{f}_j^{(0)}(x_{nj})\}^T$
2. **Cycle:** for $j = 1, \ldots, k$
   - calculate the partial residuals:

     $r_i^j = y_i - \bar{y} - \sum_{k=1}^{j-1} \hat{f}_k^{(l+1)}(x_{ik}) - \sum_{k=j+1}^{p} \hat{f}_j^{(l)}(x_{ik}), i = 1, \ldots, n$
   - calculate new $\hat{\mathbf{f}}_j$ estimates through the following expression: $\hat{\mathbf{f}}_j^{(l+1)} = \mathbf{S}_j\mathbf{r}^j$, being $\mathbf{S}_j$ the nxn smoothing matrix regarding X_j, $\mathbf{r}^j = (r_1^j, \ldots, r_n^j)^T$ the vector of partial residuals regarding X_j, $\hat{\mathbf{f}}_j^{(l+1)} = \{\hat{f}_j^{(l+1)}(x_{1j}), \ldots, \hat{f}_j^{(l+1)}(x_{nj})\}$ and l the iteration count.
3. **Repetition** of step 2 till there are no more significant differences between $\hat{\mathbf{f}}_j^{(l)}$ and $\hat{\mathbf{f}}_j^{(l+1)}$ for every j. The stoppage criterion can be defined as:

$$\frac{\sum_{i=1}^{n}(\hat{f}_j^{(l+1)}(x_{ij}) - \hat{f}_j^{(l)}(x_{ij}))^2}{\sum_{i=1}^{n}(\hat{f}_j^{(l)}(x_{ij}))^2} \leq \delta \tag{3.16}$$

   $j = 1, \ldots, p$, for any δ small enough.

Through these models it is possible to estimate covariate functional forms that reflect more accurately the true relationship between **X** and Y.

3.2.4 Generalized Additive Models

The generalized additive models are an extension of the generalized linear models, as the additive models are an extension of the linear regression models. GAMs are more flexible than additive models because the

additive predictor is linked to the response variable Y through a known link function $g(\cdot)$. This function g is chosen based on the nature of Y and on the interest of the study, like in the GLMs context.

So, the GAMs are defined as:

$$\begin{aligned} g(\mu) &= \alpha + \sum_{j=1}^{k} f_j(X_j) \\ \mu &= h(\alpha + \sum_{j=1}^{k} f_j(X_j)) \end{aligned} \tag{3.17}$$

being $\mu = E(Y|\mathbf{X})$, with Y belonging to the Exponential Family, α and $\mathbf{f}_j$ are unknown, $j = 1, \ldots, k$.

3.2.4.1 Estimation Algorithm

In this context, the partial functions $\mathbf{f}_j$ estimation is done through the *local scoring* algorithm (Hastie and Tibshirani, 1986).

This algorithm is composed of two cycles, one (the *backfitting* algorithm) nested in another, as described:

1. **External Cycle Initialization:** $\alpha^{(0)} = h^{-1}(n^{-1}\sum_{i=1}^{n} y_i)$ and $\hat{\mathbf{f}}_1^{(0)} = \cdots = \hat{\mathbf{f}}_k^{(0)} = 0$

2. **External Cycle:** For $l = 0, 1, 2, \ldots$ calculate the adjusted dependent variable:

$$z_i^{(l)} = \hat{\eta}_i^{(l)} + (y_i - \hat{\mu}_i^{(l)}) \left(\frac{\partial \hat{\mu}_l}{\partial \hat{\eta}_l}\right)_{(l)} \tag{3.18}$$

   and the weights:

$$\hat{w}_i = \left(\frac{\partial \hat{\mu}_i}{\partial \hat{\eta}_i}\right)^2_{(l)} (\hat{V}_i^{(l)})^{-1} \tag{3.19}$$

   , being $\hat{\eta}_i^{(l)} = \hat{\alpha}^{(l)} + \sum_{j=1}^{p} \hat{f}_j^{(l)}(x_{ij})$, $\hat{\mu}_i^{(l)} = h(\hat{\eta}_i^{(l)})$, $\left(\frac{\partial \hat{\mu}_i}{\partial \hat{\eta}_i}\right)_{(l)} = \frac{\partial h}{\partial \eta}\Big|_{\eta=\hat{\eta}_i^{(l)}}$ and $\hat{V}_i^{(l)}$ represents the variance of Y calculated on $\hat{\mu}_i^{(l)}$, $i = 1, \ldots, n$.

3. **Internal Cycle** (*Backfitting* for $\hat{\mathbf{f}}^{(l+1)}$ estimation): Adjust an additive model to $\mathbf{z} = (z_1, \ldots, z_n)$, weighted by $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_n)$, in order to get $\hat{\mathbf{f}}_j^{(l+1)}$, $\hat{\eta}_i^{(l+1)}$ and $\hat{\mu}_i^{(l+1)}$, with $j = 1, \ldots, k$ and $i = 1, \ldots, n$.

4. **Repeat the steps 2 and 3** replacing $\hat{\mathbf{f}}_j^{(l)}$, $\hat{\eta}_i^{(l)}$, $\hat{\mu}_i^{(l)}$ for $\hat{\mathbf{f}}_j^{(l+1)}$, $\hat{\eta}_i^{(l+1)}$, $\hat{\mu}_i^{(l+1)}$ till the stoppage criterion is met, for example:

$$\frac{\sum_{i=1}^{n}(\hat{\eta}_i^{(l+1)} - \hat{\eta}_i^{(l)})^2}{\sum_{i=1}^{n}(\hat{\eta}_i^{(l)})^2} \leq \delta \tag{3.20}$$

   for any δ small enough.

It is noteworthy that the *local scoring* algorithm is not restricted to the *backfitting* in its internal cycle. Any iterative process capable of solving equation 3.15 can be introduced.

3.3 Some Basis in Epidemiology and Propensity Score Methodology

3.3.1 The Internal Validity of Observational Studies

In 1957, Campbell developed the internal validity concept which consists on 3 principles, that serve to validate any inference relative to a cause-effect relationship (Campbell, 1957). Being C the cause and E the effect, these are the principles:

- C must preceed E in time;
- C covaries with E;
- There is no other plausible explanation for this covariation.

Shadish et al identified many threats to internal validity (Shadish et al., 2002). But, the major threat we face on observational studies is the selection threat. The selection threat is the risk of selecting patients to treatment and control group, which are significantly different relative to their observable and unobservable characteristics, in a way that bias the treatment effect estimate (selection bias) (Guo and Fraser, 2009).

3.3.2 Confounding

In observational studies, the groups assigned to treatment or control are, naturally, imbalanced with respect to many covariates. This is because these studies have to perform the analysis with pre-established treatment groups where the treatment assignment is not randomized. In contrast, experimental studies compare treatment groups where the distribution of certain/all variables is balanced. Here, the treatment assignment is controlled and performed in order to assure the balance of covariates that potentially influence the treatment effect estimate (potential confounders) between groups, through multiple randomization methods (D 'agostino, 1998).

If one proceeds to estimate the treatment effect by directly comparing the treatment and the control group, in an observational study, there is the risk to obtain a confounded/biased estimate (D 'agostino, 1998). This is due to covariates that have different distributions between treatment groups and, simultaneously, have an association with the outcome (confounding variables). So, there is the need to control/balance these type of variables, and therefore eliminate the selection bias/confoundness. Basically, confounding is a distortion of an association between an exposure and an outcome brought about by extraneous/confounding factors (Gordis, 2014).
Because of all this, the experimental studies are considered the "gold standard", and the observational ones are often performed when there is lack of funding or when the experimental studies are not ethical or simply infeasible (Meldrum, 2000).

3.3.3 Neyman-Rubin's Counterfactual Framework

The intuitive idea of comparing two exactly equal populations relative to their characteristics, and assign to one group the treatment and another group the control (the counterfactual), is to estimate an unbiased treatment effect. Of course, this is often infeasible. As for each treated subject is very difficult to find other subject in the control group with exactly the same characteristics (the counterfactual). This is a missing data problem because it is unknown what would be the outcome of the treated subject if he was assigned to control, and vice-versa (Austin, 2011a).

The Neyman-Rubin's counterfactual framework formalizes the idea that each individual has two potential outcomes. The potential outcome if treated and the potential outcome if not treated, and only one is observable (Guo and Fraser, 2009). This is because one subject cannot belong to the treatment and to the control group, simultaneously. The equation 3.21 describes the basis of the Neyman-Rubin's Counterfactual Framework (NRCF).

$$Y_i = Z_i Y_i(1) - (1 - Z_i) Y_i(0) \tag{3.21}$$

Where Z_i is the dummy variable which represents the treatment received by subject *i* (1 or 0) , $Y_i(1)$ the potential outcome if the subject is treated ($Z_i = 1$) and $Y_i(0)$ the potential outcome if the subject is not treated ($Z_i = 0$). Y_i is the "observed" outcome for subject i, i.e., it is the outcome that results from the treatment assignment to subject i. In this context, Z_i functions as a switch variable because, if $Z_i = 1 \Rightarrow Y_i = Y_i(1)$ and if $Z_i = 0 \Rightarrow Y_i = Y_i(0)$.

As only one outcome is observable, the other one is always missing. This is the missing data problem stated before. In order to estimate the missing value, one could find a subject with the exact same **x** who was assigned to the opposite treatment assignment condition and measure its outcome, but this is often infeasible.

This missing outcome value can be estimated too by using a PS approach. Based on the properties of PS, it is possible to select subjects which, assymptotically, have the same distribution of **X**. The outcomes of these subjects are a great estimate of the unobserved potential outcome. This way, an unbiased treatment effect estimate can be estimated.

3.3.3.1 Assumptions

As the context of this work will be within the NRCF, there are two assumptions that need to hold. To prevent biased estimates due to unmeasured effects, Rosenbaum and Rubin proposed:

Ignorable Treatment Assignment Assumption (ITAA): This assumption, also known as the *unconfoundness assumption* or the *no unmeasured confounders assumption* (Rubin and Rosenbaum, 1983), holds when the treatment assignment (Z) is independent from the potential outcomes $\{Y(1), Y(0)\}$, if the observable covariates X are held constant. ITAA is represented as:

$$\{Y(1), Y(0)\} \perp Z|\mathbf{X=x} \tag{3.22}$$

for all x where $\perp$ means statistical independence.

In order to get an unconfounded estimate, the ITAA has to hold conditional on the observed covariates. This assumption is violated if there are confounding variables that were not observed, and therefore will provoke an hidden bias in the final estimate.

"*This condition states that, for those having $X = x$, the assignment rule is determined by an independent Bernoulli random variable having a probability of success $Pr(Z = 1|X = x)$*" (Emura et al., 2008)

In Randomized controlled trials (RCT), this assumption holds naturally, because the randomization balances the covariates **X** between treated and control groups. In this context, the treatment assignment is independent from any covariate.

Stable Unit Treatment Value Assumption (SUTVA) This assumption holds if the outcome variable is independent from the mechanism used in treatment assignment, and from the treatments other units receive.

3.3.4 Types of Treatment Effects

The treatment effect for each subject i is defined as $Y_i(1) - Y_i(0)$. The **Average Treatment Effect (ATE)** (denoted by τ_{ATE}) is given by $E[Y(1)] - E[Y(0)]$, and represents *"the average effect, at the population level, of moving an entire population from untreated to treated."* (Austin, 2011a). If $y_i(1)$ and $y_i(0)$ were observable for every subject, an estimate would be: $\hat{\tau}_{ATE} = \frac{1}{n}\sum_{i=1}^{n} y_i(1) - y_i(0)$.

Other type of treatment effect is the **Average Treatment Effect for the Treated (ATT)** (denoted by τ_{ATT}) which is given by $E(Y(1)|Z = 1) - E(Y(0)|Z = 1)$ and represents the average effect of treatment on those subjects who were treated. If $y_i(0)$ was observable for all treated subjects n_1, an estimate would be $\hat{\tau}_{ATT} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i(1) - y_i(0)$. ATT is particularly useful in policy contexts where the interest is not whether the treatment is effective for all population, but if it is effective for those subjects who are targeted to be treated. In a randomized study, these two measures coincide because the treated population will not differ from the overall population.

These treatment effects are marginal effects, i,e. they measure the average treatment effect on the population. These are different from the **Conditional effects** that arise when performing any kind of regression, and, usually, are represented by the β coefficients.
"A marginal effect is the average effect of treatment on the population. A conditional effect is the average effect of treatment on the individual" (Austin, 2011a).

3.4 The Propensity Score

In 1983, Rosenbaum and Rubin proposed an alternative and innovative method to estimate the treatment effect in observational studies, removing the effects of confounding: the propensity score analysis.

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Assuming $e(\mathbf{X})$ the propensity score, Z the treatment assignment variable and $\mathbf{X}$ the covariates:

$$e(\mathbf{X}) = Pr(Z = 1|\mathbf{X}). \tag{3.23}$$

The unique properties of the propensity score (PS) allow the treatment effect estimation, in multiple ways (described further). PS is a balancing score, which means that the distribution of $\mathbf{X}$ is similar in both treated and control groups, conditionally on $e(\mathbf{X})$. This large sample property is defined as:

$$\mathbf{X} \perp Z|e(\mathbf{X}). \tag{3.24}$$

It means that if two populations (treated and non-treated) have the same value of e($\mathbf{X}$), the outcome mean difference between them is an unbiased estimate of the treatment effect, if $\mathbf{X}$ contains all the confounding variables (i.e. if ITAA holds). Populations with the same e($\mathbf{X}$) are comparable because they have the same distribution regarding potential confounding variables. This is why PS analysis has the ability to, partly, mimic a randomized study.

This balancing property allows to replace **X** by **e(X)** in the analysis, which is particularly useful when **X** is of high dimension. Because of this, e(**X**) can be used in matching, stratification and LR replacing **X**.

"The Propensity score $e(\boldsymbol{x}_i)$ is a balancing measure (so called the coarsest score) that summarizes the information of vector $\boldsymbol{x}_i$ in which each x is the finest score." (Guo and Fraser, 2009).

Propensity Score Model The first step in every propensity score analysis, is to estimate the propensity score for every subject $\hat{e}(\mathbf{x}_i)$. This is, usually, done through a logistic regression, where the treatment assignment variable (Z) is the dependent variable and **X** the independent variables. The fitted values of this LR are the estimated propensity scores. The selection process by which the independent variables are selected to be part of the model is discussed further.

$$e(\mathbf{X}) = Pr(Z = 1|\mathbf{X}) = \frac{exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)} \tag{3.25}$$

3.5 Propensity Score Matching

3.5.1 Traditional Matching

The objective of this section is to introduce the concept of matching. The first matchings consisted on pairing each treated subject with other subjects who were not treated, with similar specified observed covariates **x**. This covariates are the ones that are unbalanced between groups and are potential confounding variables. Consider y_{j1} and y_{j0} the observed outcomes of the j^{th} pair ($j = 1, \ldots, n$), and $D_j = Y_{j1} - Y_{j0}$. The treatment effect (τ) estimate can be obtained like this:

$$\hat{\tau} = \bar{D} = \frac{1}{n} \sum_j (y_{j1} - y_{j0}) = \bar{y}_1 - \bar{y}_0. \tag{3.26}$$

If the outcome is continuous, $\hat{\tau}$ is the estimated mean outcome difference between groups. If the outcome is dichotomous, $\hat{\tau}$ is the estimated difference between the proportion of subjects experiencing the outcome event in each group.

As the estimate derives from a paired experiment, the variance of the estimate has to be calculated accordingly . For a continuous outcome, can be used a paired *t-test*, and for a dichotomous one can be used the McNemar's Test (Imbens, 2004).

By pairing subjects whose values on the confounding variables are similar, a greater fraction of the difference in the outcome (D_j), is due to the treatment assignment, and a lesser fraction is due to the confounder (Rubin, 1973) .

If there are many covariates to match on, it is possible few pairs being matched, which jeopardizes the significance of the estimate. For example, suppose there are k dichotomous covariates to match on, it means that exist 2^k possible values of **x**. This problem is called the *dimensionality of matching*.

3.5.2 Matching on the Propensity Score

Basically the same as matching on **X**. But, instead of pairing based on the covariates **X**, the pairing is done through $\hat{e}(\mathbf{x})$. Treated subjects are paired with non-treated subjects, if both have similar values of $\hat{e}(\mathbf{x})$. This is done because subjects with the same $\hat{e}(\mathbf{x})$, have, asymptotically, the same multivariate distribution of **X** and, therefore, are comparable (Remind the large sample property on Equation 3.24). This leads back to the Neyman-Rubin Counterfactual Framework because matching is all about finding

counterfactuals for the treated subjects, so that one can compare the outcome of a treated subject $Y_i(1)$ to the outcome of its counterfactual $\hat{Y}_i(0)$, and therefore obtain a treatment effect estimate for the treated $\tau_{i,ATT} = Y_i(1) - \hat{Y}_i(0)|Z_i = 1$. That is why, this work is within the NRFC.

So, if the matching is done based on $\hat{e}(\mathbf{x})$, the two matched samples will share the same distribution of $\mathbf{X}$. This matching approach address the *dimensionality of matching* problem.

In practice, the matching is not done on $\hat{e}(\mathbf{x})$, but on its tranformation (i.e., $\hat{q}(\mathbf{x}) = log[(1-\hat{e}(\mathbf{x}))/\hat{e}(\mathbf{x})]$). This way, the $\hat{q}(\mathbf{x})$ values are not compressed between 0 and 1 like $\hat{e}(\mathbf{x})$ and are more likely to be normal distributed which is an advantage in genetic matching (described further) (Rosenbaum and Rubin, 1985).

There are several choices to be made when applying matching:

- **Greedy or Optimal** In greedy matching, at first, a treated subject is chosen at random. Then, this treated subject is matched with an untreated subject whose $\hat{e}(\mathbf{x})$(PS) is the closest to the PS of the treated subject. This process is repeated until all the treated subjects are matched (if there is a suitable match for every treated subject). This is called "greedy" because it does not minimize the total within-pair PS difference , contrary to optimal matching. It was proved that both methods have the same performance in the production of balanced matched samples (Austin, 2011a).

  The most common algorithm to match subjects on the PS is a greedy one called ***Nearest Neighbour Matching with a Specified Caliper Distance***(NNM-CD). This algorithm is very simple and efficient. Basically, it states that each treated subject is matched with the untreated subject whose PS is the closest, but this difference cannot exceed a pre-specified threshold. This limit is called *caliper distance*. If a given treated subject have no untreated subjects whose PS lays within the caliper distance, this subject remains unmatched and is excluded from the analysis. Austin suggested to use a caliper distance that is 0.2 of the pooled (treatment and control groups) standard deviation of the PS estimates. (Austin, 2011b).

- **Replacement or No Replacement** If the option is matching without replacement, every untreated subject that is matched with a treated subject, is no longer available for matching thereafter. On the other hand, if matching with replacement is chosen, an untreated subject can be matched with multiple treated subjects. Matching with replacement produces the most balanced samples and the lowest conditional bias. *Matching with replacement produces matches of higher quality than matching without replacement by increasing the set of possible matches* (Abadie and Imbens, 2006).

Common Support Region Problem In order to match a substantial number of pairs, it is needed that the two groups (treated and non-treated) have similar estimated propensity scores distributions. If it is not the case, at least, the two distributions have to share a common support region to let a sufficient number of pairs to be formed. This is called the *overlap assumption*.

For example, in the figure 3.1, the *overlap assumption* is not violated because there still is a significant common support region. If this assumption is violated, matching estimators are not appropriate.

3.5.3 Variable Selection

The covariates chosen to be introduced in the PS model (Equation 3.25), have to be selected in order to find the PS estimates that most balance the matched treatment groups, regarding the covariates that

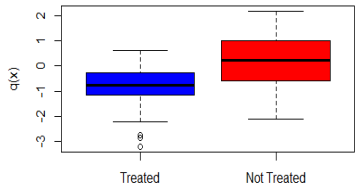


Figure 3.1: Example: Estimated propensity scores $\hat{q}(x)$ distributions in treated and non-treated groups. Represented by a box-plot.

influence the outcome. The variables introduced in the model are the ones that will be controlled/balanced between the matched treatment groups, as stated by the balancing property $\mathbf{X} \perp \mathbf{Z}|e(\mathbf{X})$. Thus, it would be obvious to include only the variables that are prognostically important, but the identification of these variables is not straightforward. Because of this, there are four possible variable subsets that are commonly tested within the PS model:

- True Confounders (variables associated with both treatment assignment and the outcome);
- Potential Confounders (variables associated with the outcome);
- Variables associated with the treatment assignment;
- All observed variables.

The introduction of potential confounders or true confounders in the model proved to generate more precise treatment effect estimates than subsets 3 and 4. This makes sense because, in this case, the procedure is focused in balancing only these variables. Although, if there are unidentified confounding variables, the estimate will be biased. Including all observed covariates in the model, i.e. balancing all observed variables between the matched treatment groups, is reassuring because the identification of confounders can be tricky. Other option is to include only the variables that are associated with the treatment assignment, assuming that all the other variables will be balanced after matching because they do not influence the treatment assignment. In simulation studies, all subsets produce equally biased estimates (Austin et al., 2007).

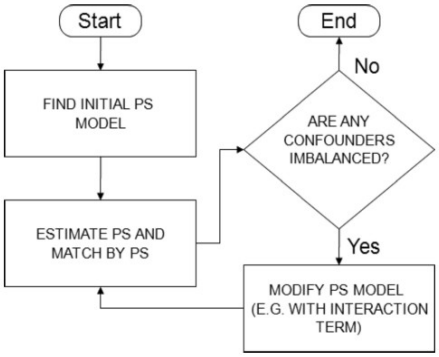


Figure 3.2: Search for the best Propensity Score model. Figure taken from Diamond and Sekhon (2012).

Treatment and control subjects are matched on PS, in order to obtain two comparable populations with the same $\mathbf{X}$ distribution of interest, and therefore to estimate the treatment effect. If the matched

samples are balanced regarding this **X** variables included in the model, it means the propensity score model serves our best interest and the matching was successful. If the matched groups are not balanced, there is the need to reformulate the PS model (e.g. by including/excluding a square term of a covariate or an interaction term) (Figure 3.2). Insights on variable balance within the matched samples will be presented further.

In this context, discriminative and predictive power measures are not sufficient to access the PS model correctness (Stürmer et al., 2006). This because, our main objective is to balance the treatment groups regarding the covariates within the PS model and not to predict the treatment assignment probability. *"Since the propensity score ($e(\boldsymbol{X})$) is a balancing score ($Z \perp \boldsymbol{X}|e(\boldsymbol{X})$), the estimate of the propensity score is consistent only if matching on this propensity score asymptomatically balances the observed covariates."* (Diamond and Sekhon, 2012)

3.5.4 Balancing Diagnostics

To assure that matching was successful, one needs to check if both groups have the same distribution regarding the variables within the PS model. If this happens to be, the PS estimates are consistent because the PS model is well specified, if not, the PS model needs to be changed in order to improve balance.

The best way to access the covariates balance between groups. One of them is the Standardized Mean Difference (SMD) . This measure is defined as

$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

for continuous variables, and

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{(\hat{p}_{treatment})(1-\hat{p}_{treatment}) + \hat{p}_{control}(1-\hat{p}_{control})}{2}}}$$

for dichotomous variables, where $\bar{x}$ denote the group mean for a continuous variable, s^2 the group sample variance and $\hat{p}$ the group proportion for a dichotomous variable. If d is less than 0.1, the difference between groups regarding that specific variable is considered negligible (i.e. the null hypothesis H_0 is not rejected for $H_0 : \mu_T = \mu_C$) (Normand et al., 2001).

Significance tests like the T-test, Mann-Whitney or Kolmogorov Smirnov should not be used in these circumstances because the significance levels can be confounded with sample size (Flury and Riedwyl, 1986; Austin, 2008). This means that when the sample size is very large, the differences are very likely to be statistically significant. Other way to analyse the balance is to apply graphical methods like side-by-side box plots, quantile-quantile plots or cumulative distribution functions.

3.5.5 Genetic Matching

This section is about the Genetic Matching algorithm published by Diamond and Sekhon (2012), and implemented by the R package *Matching*.

The genetic matching is a process that iteratively modifies the matching metric to achieve a maximum covariate balance between groups.

In the previous version of matching, subjects were matched simply on their PS. In genetic matching, subjects are matched on the Generalized Mahalabonis Distance (GMD) between them. But, first of all, we must review the concept of Mahalanobis Distance (MD).

Mahalanobis Distance MD is a multivariate distance measure between two subjects. MD is unitless and scale-invariant, and takes into account the correlations between variables. For example, MD between the subject i and j, where $\boldsymbol{X}$ is the covariate matrix (nxk) and $\boldsymbol{S}$ is the sample covariate matrix (nxn) of $\boldsymbol{X}$, is represented this way:

$$MD(X_i, X_j) = \sqrt{(X_i - X_j)^T \boldsymbol{S}^{-1} (X_i - X_j)}.$$

This measure can be used to match the subjects, but it has a low performance when covariates have non-ellipsoidal distributions. This is why Rosenbaum and Rubin (1985) proposed to include the PS estimates among the other covariates in **X**. Thus, all the covariates are represented through the PS estimates.

Generalized Mahalanobis Distance(GMD) GMD is very similar to MD. The only difference is that GMD assigns weights ($\boldsymbol{W}$) to all matching variables. These weights are assigned by the algorithm and reflect the importance of each variable to achieve the best overall balance. Considering $\boldsymbol{W}$ the weight matrix (kxk), and considering that $\boldsymbol{W}$ is restricted to 0 except for the main diagonal, the GMD between two subjects i and j is:

$$GMD(X_i, X_j, \boldsymbol{W}) = \sqrt{(X_i - X_j)^T (\boldsymbol{S}^{-1/2})^T \boldsymbol{W} \boldsymbol{S}^{-1/2} (X_i - X_j)}.$$

If the maximum balance is achieved by matching on the PS, the W matrix will have all its values restricted to 0 except the one regarding the PS covariate. If the maximum balance is achieved by matching on the MD, the main diagonal will have values greater than 0, except for the PS covariate value.

Loss Function There is the need to choose which statistic will be used to access the balance. The most popular choices are the standardized mean difference and the *KS* statistic. Examples of loss functions are the maximum value of the KS statistic or the maximum standardized mean difference across all variables. If the loss function is lower, the sample is better balanced on **X**. Otherwise, if it is higher, the balance is worse. The all point of genetic matching is to search, iteratively, for a specific $\boldsymbol{W}$ matrix that minimizes the loss function.

Iterative Algorithm The genetic matching search for the ideal $\boldsymbol{W}$ matrix. First, it generates a batch of $\boldsymbol{W}$ matrixes. Next, perform as many matchings as the number of $\boldsymbol{W}$ matrixes. Then, compute the loss function for every matched sample. The matched sample with the lowest loss function value is selected, and its W matrix will be used to generate a new similar batch of W matrixes. This algorithm moves toward the minimum loss function value possible. The algorithm stops running when the maximum number of iterations is reached. This procedure is depicted in figure 3.3.

This type of matching cannot be used to access the PS model correctness, because the PS estimated values are not used directly on matching. Therefore, the balancing score property ($Z \perp \boldsymbol{X} | e(\boldsymbol{X})$) is not available for checking.

To conclude, GM tries to search for the GMD metric which provides the most balanced matched sample. Basically, this algorithm matches the subjects, evaluate the balance and continues to learn iteratively. In this context, one still needs to opt between matching with or without replacement, and between greedy and optimal.

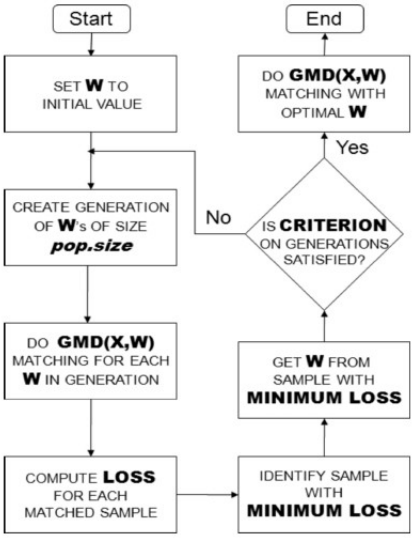


Figure 3.3: Representation of Genetic Matching. Figure Taken from Diamond and Sekhon (2012)

3.5.6 Estimation of the Average Treatment Effect for the Treated

To perform matching one have to select subjects from the non-treated group that are the counterfactuals, i.e. comparable regarding some covariates, to every subject in the treated group. These counterfactuals outcomes are an estimate of the potential outcome if the treated have not been treated ($\hat{Y}_i(0)$). This way, matching is trying to estimate the Average Treatment Effect for the Treated (ATT), and not for all the population (Deb et al., 2015). ATT is calculated as the mean difference between treated and control matched subjects, for each i^{th} treated subject :

$$\hat{\tau}_{ATT} = \frac{1}{n}\sum_i Y_i(1) - \hat{Y}_i(0) \tag{3.27}$$

Abadie and Imbens (2002) developed a variance estimator for $\hat{\tau}_{ATT}$. This estimator, under homoscedasticity conditions: 1) $\tau_i = Y_i(1) - Y_i(0)$ is constant and 2) the conditional variance of $Y_i(Z)$ given X_i does not vary with the treatment Z, is given below.

This variance estimator needs to account if the matching is done with replacement. Because of that, the expression that calculates the variance is complex. Let $J_M(l)$ denote the set of indices relative to the subjects that are matched to unit l, and $\#J_M(l)$ the number of indices within the set. Let $K_M(i)$ be the total number of times i is used as a match for all observations l of the opposite treatment condition, each time weighted by the total number of matches for observation l: $K_M(i) = \sum_{l=1}^{N} 1\{i \epsilon J_M(l)\}\frac{1}{\#J_M(l)}$, being $1\{\bullet\}$ the indicator function. And, $K'_M(i) = \sum_{l=1}^{N} 1\{i \epsilon J_M(l)\}(\frac{1}{\#J_M(l)})^2$.

The estimator variance is:

$$\begin{aligned} Var(\hat{\tau}_{ATT}) = &\frac{1}{N_1^2}\sum_{i=1}^{N}[Z_i(Y_{i1} - Y_{i0} - \tau_{ATT})^2 + (1 - Z_i)(K_M^2(i) - K'_M(i)) \\ &\frac{1}{2N_1}\sum_{i:Z_i=0}\{\frac{1}{\#J_M(i)}\sum_{l \epsilon J_M(i)}(Y_l - Y_i - \hat{\tau}_{ATT})^2\}] \end{aligned} \tag{3.28}$$

where N_1 is the number of subjects treated, and N the total number of subjects. This variance can be used to perform a significance test or to calculate a confidence interval, because $\frac{\hat{\tau}_{ATT} - \tau_{ATT}}{\sqrt{Var(\hat{\tau}_{ATT})}} \sim N(0,1)$ allowing a z test. Being the CI given by:

$$\{\hat{\tau}_{ATT} \pm z_{1-\alpha/2} \times \sqrt{Var(\hat{\tau}_{ATT})}\}. \tag{3.29}$$

3.6 Stratification

Stratification is a Matching generalization. Instead of pairing 1 on 1 subjects, stratification consists on creating groups (strata) which the subjects within each stratum have, roughly, the same PS estimate. To create these strata, PS estimates quantiles (typically quintiles) are often used as stratum delimiters. If the stratification is performed on 5 strata, quintiles are used, if performed on 10 strata, deciles are used. This stratification method is based on equal-frequency strata, but other stratification methods are valid as well (Huppler Hullsiek and Louis, 2002).

Stratification takes advantage of the PS balancing property, as matching does. Here, any subject (treated or non-treated) within the same stratum has, roughly, the same **X** distribution. This makes the two treatment groups comparable within each stratum (Rosenbaum and Rubin, 1984).

The utility of Stratification has been proved a long time ago. Stratifying (using quintiles) on one confounding variable (not on PS estimates) was proved to reduce the confounding by 90% (Cochran, 1968). In 1984, Rosenbaum and Rubin extended this result to the PS stratification, which reduces the effects of confounding due to measured confounders by, approximately, 90% as well (Rosenbaum and Rubin, 1984). As the number of strata increases, the bias reduction is improved, but the ATE estimate variance increases (Huppler Hullsiek and Louis, 2002).

3.6.1 Estimation of the Average Treatment Effect

Here the ATE (Average Treatment Effect) is estimated this way:

$$\hat{\tau}_{ATE} = \frac{1}{K}\sum_{j=1}^{K}\hat{\tau}_{ATE_j} = \frac{1}{K}\sum_{j=1}^{K}\bar{y}_{1j} - \bar{y}_{0j}, \tag{3.30}$$

being K the number of strata, $\bar{y}_{1j}$ the mean outcome of the treated subjects in stratum j, $\bar{y}_{0j}$ the mean outcome of the non-treated subjects in stratum j, and $\hat{\tau}_{ATE_j}$ the stratum specific estimate.

The $\hat{\tau}_{ATE}$ estimator variance is estimated this way:

$$Var(\hat{\tau}_{ATE}) = \frac{1}{K^2}\sum_{j=1}^{K}Var(\hat{\tau}_{ATE_j}) = \frac{1}{K^2}\sum_{j=1}^{K}\hat{\sigma}_j^2, \tag{3.31}$$

being the stratum-specific variance estimate $\hat{\sigma}_j^2$:

$$\hat{\sigma}_j^2 = \frac{1}{n_{1j}}s_{1j}^2 + \frac{1}{n_{0j}}s_{0j}^2 \tag{3.32}$$

with

$$\begin{aligned} s_{1j}^2 &= \frac{1}{n_{1j}}\sum_{i=1}^{n_j}(Z_iY_i - \bar{y}_{1j})^2 \qquad & s_{0j}^2 &= \frac{1}{n_{0j}}\sum_{i=1}^{n_j}\{(1-Z_i)Y_i - \bar{y}_{0j})\}^2 \\ \bar{y}_{1j} &= \frac{1}{n_{1j}}\sum_{i=1}^{n_j}Z_iY_i & \bar{y}_{0j} &= \frac{1}{n_{0j}}\sum_{i=1}^{n_j}(1-Z_i)Y_i^2 \end{aligned} \tag{3.33}$$

n_j the number of subjects in stratum j, n_{1j} the number of treated subjects in stratum j, n_{0j} the number of non-treated subjects in stratum j. Z_i and Y_i, the treatment assignment and the outcome for each subject i.

As $\frac{\hat{\tau}_{ATE}-\tau_{ATE}}{\sqrt{Var(\hat{\tau}_{ATE})}} \sim N(0,1)$, so a z test is possible. Being the CI given by:

$$\{\hat{\tau}_{ATE} \pm z_{1-\alpha/2} \times \sqrt{Var(\hat{\tau}_{ATE})}\}. \tag{3.34}$$

3.7 Inverse Probability of Treatment Weighting

Inverse Probability of Treatment Weighting (IPTW) using PS is a weighting technique that creates a synthetic sample in which the distribution of measured confounders is independent of the treatment assignment (mimicking this way a randomized clinical trial). IPTW consist on weighting more the subjects that represent populations that are less represented within a specific treatment group. When the word "population" appears in this context, it means: subjects with roughly the same PS and therefore with, asymptotically, the same **X**.

It is natural that subjects with higher PS estimates are more likely to receive treatment, and therefore, the most represented populations in the treated group are the ones that correspond to higher PS and vice-versa. To achieve synthetic samples that represent the all population, the strategy is to weight more the subjects with lower PS estimates in the treated group and the subjects with higher PS estimates in the non-treated group. These weights are the inverse of the probability of receiving the treatment actually received. So, for each subject the weight is:

$$w_i = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i}, \tag{3.35}$$

being Z_i the treatment assignment and e_i the PS for subject i (Austin, 2011a). The sum of weights is roughly equal in both treatment groups, although the number of subjects in each group may be very different. For the treated subjects ($i = 1, \ldots, n_1$), $\sum_{i=1}^{n_1} w_i = \sum_{i=1}^{n_1} \frac{1}{e_i} \approx n_0 + n_1 = n$ and for the non-treated subjects ($j = 1, \ldots, n_0$), $\sum_{j=1}^{n_0} w_j = \sum_{j=1}^{n_0} \frac{1}{1-e_j} \approx n_0 + n_1 = n$. This means that the new weighted (synthetic) samples have the same dimension (the total number of treated and non-treated n). These new samples represent the all range of populations (i.e. the all range of the PS estimates), and therefore are great to estimate the ATE.

3.7.1 Estimation of the Average Treatment Effect

The ATE is estimated by a weighted mean, using the weights referred before:

$$\begin{aligned}
\hat{\tau}_{ATE} &= \frac{1}{n}\sum_{i=1}^{n_1} Y_i w_i - \frac{1}{n}\sum_{j=1}^{n_0} Y_j w_j \\
&= \frac{1}{n}\sum_{i=1}^{n_1} Y_i \frac{1}{\hat{e}_i} - \frac{1}{n}\sum_{j=1}^{n_0} Y_j \frac{1}{1 - \hat{e}_j} \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{Z_i}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n} Y_i \frac{1 - Z_i}{1 - \hat{e}_i}.
\end{aligned} \tag{3.36}$$

The variance of $\hat{\tau}_{ATE}$ estimator is estimated this way:

$$Var(\hat{\tau}_{ATE}) = \frac{1}{n^2}\sum_{i=1}^{n} \hat{I}_i^2, \tag{3.37}$$

being

$$
\begin{aligned}
\hat{I}_i^2 &= \frac{Z_i Y_i}{\hat{e}_i} - \frac{(1-Z_i)Y_i}{1-\widehat{e}_i} - \widehat{\tau}_{ATE} - (Z_i - \hat{e}_i)\widehat{\mathbf{H}}_{\beta,1}^T \widehat{\mathbf{E}}_{\beta\beta}^{-1}\mathbf{W}_i \\
\widehat{\mathbf{H}}_{\beta,1}^T &= \frac{1}{n}\sum_{i=1}^{n}\{\frac{Z_i Y_i(1-\hat{e}_i)}{\hat{e}_i} + \frac{(1-Z_i)Y_i\hat{e}_i}{1-\hat{e}_i}\}\mathbf{W}_i \\
\widehat{\mathbf{E}}_{\beta\beta}^{-1} &= \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i(1-\hat{e}_i)\mathbf{W}_i\mathbf{W}_i^T \\
\hat{e}_i &= 1/\{1+exp(-\mathbf{W}_i^T\boldsymbol{\beta})\} \\
\mathbf{W} &= \mathbf{X}^T \qquad \text{(if PS estimated by LR)}
\end{aligned}
\tag{3.38}
$$

A z test is possible to compute because $\frac{\hat{\tau}_{ATE}-\tau_{ATE}}{Var(\hat{\tau}_{ATE})} \sim N(0,1)$.

3.8 GLM and PS mixed approach: Covariate Adjustment

As seen before, the PS has the ability to replace the vector of covariates **X** in multiple analysis. Thus, using the PS as a covariate in a GLM is also an option.

The model proposed here consists on the outcome (Y) as the dependent variable, and as independent variables:

- the treatment assignment (Z);
- Prognostically Important variables $(X_1, X_2, \ldots, X_p)$;
- the PS estimated based on all the other variables ($e(X_{p+1}, X_{p+2}, \ldots, X_k)$).

The prognostically important variables are described by Hirano and Imbens as the variables that are worth to be included in the model individually, i.e., the variables that are associated with the outcome (Hirano and Imbens, 2001). The PS estimated based on all the other observed variables are like a complement to the model. The term $e(X_{p+1}, X_{p+2}, \ldots, X_k)$ is seen as a summary of all "non-important" variables that are not relevant enough to be individually on the model.

As in the context of this work, the outcome variable is dichotomous, so, a logit link will be used. The β coefficients estimation is the same as in 3.1. The referred model is given by:

$$
log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 Z + \beta_2 X_1 + \beta_3 X_2 + \cdots + \beta_{p+1}X_p + \beta_{p+2}e(X_{p+1}, X_{p+2}, \ldots, X_k). \tag{3.39}
$$

This approach is an attempt to build a more parsimonious model. This is, particularly, useful when there are large number of variables to control.

3.9 Comparison of all Propensity Score Methods

Table 3.4 summarizes advantages and disadvantages of each PS method described before.

Table 3.4: Advantages and Disadvantages of all PS methods and their respective Treatment Effect Estimated. Table retrieved, partly, from Deb et al. (2015)

Methods	Advantages	Disadvantages	Treatment Effect Estimated
Matching	• Superior at reducing bias compared with stratification and covariate adjustment	• Unmatched treated participants, and possibly unmatched control participants, are excluded from the analysis, decreasing precision of the treatment effect estimate and external validity • Requires more control participants than treated participants	ATT
Stratification	• Uses all Data	• Reduces bias less than matching and IPTW	ATE
IPTW	• Superior at reducing bias compared with stratification and covariate adjustment • Uses all Data	• May be more sensitive to mis-specification of the propensity-score model and extreme propensity-score values	ATE
Cov. Adjust.	• Uses all Data	• No clear distinction between the design phase and the analysis phase • Very Biased estimates	OR

3.10 Comparison between Propensity Score and Regression Methods

Historically regression methods have been more used than Propensity Scores (PS) when analysing observational data (Austin et al., 2007). This happens because regression methods allow to identify indivually the factors responsible for the outcome, something that PS analysis is unable. However, Propensity Scores are increasingly being used as a strategy to control confoundness (Arbogast and VanderWeele, 2013), and the reasons why are explained right away.

In the cardiology context, tendentiously there are many variables to control, which can be problematic for regression methods to succeed. Previous research suggested a minimum of 10 events for each covariate included in the regression model (Peduzzi et al., 1996). In PS analysis, even if the event of interest is rare and treatment/non-treatment is common, one can control for all baseline variables successfully, contrary to what happens in regression methods (Braitman and Rosenbaum, 2002). This because the probability of treatment assignment (i.e. the propensity score) is estimated in good conditions according to Peduzzi et al. (1996) (1 covariate for each 10 events), even if the events are rare. This is the main reason why Propensity Scores were used in this work, because the events are rare (27) and the propensity scores can be estimated in good conditions by logistic regression or generalized addictive models.

Additionally, PS analysis has advantages like the separation between design phase (when Propensity Score estimates are obtained) and analysis phase. In regression methods, the treatment effect estimate is always observable while building the model, thus this can be a temptation to continually modify the model until the desired association is achieved. In PS analysis, only after checking the balance of the baseline covariates (design phase) the analysis is performed.

Another great advantage of PS analysis is that one can, explicitly, compares two treatment groups regarding the observed baseline covariates. If the two populations are not comparable regarding the prognostically important variables (i.e. if there is no overlap), the PS analysis is not correct. This is an intuitive thought too, if two populations are not comparable any treatment effect estimation is possible. Using regression methods, it is difficult to access this overlap. So, if one tries to apply a regression-based approach to treatment groups with no overlap, an interpolation is being made between two different populations (Austin, 2011a).

Chapter 4

Results

4.1 Logistic Regression

4.1.1 Univariate Analysis

First of all, an univariate logistic regression (LR) is performed to find the variables that are useful to explain the outcome variable. All variables that have a p-value lower than 0.25 in its univariate logistic regression, are selected to be part of the preliminary model. The descriptive analysis and the corresponding univariate p-values for every variable observed are shown in Table 4.1.

Table 4.1: Contingency tables for every categorical variable in the dataset regarding each study endpoint group. Average and Standard Deviation(SD) for continuous variables with normal distribution and Minimum(min.)/Median/Maximum(max.) for non-normal continuous variables, regarding each study endpoint group. p-value obtained from the univariate logistic regression (LR).

Variable	Variable Values		Non-occurrence of Stroke or TIA (n=16683)	Occurrence of Stroke or TIA (n=27)	Total	LR p-value
Demographic						
Age	QC	Average (SD)	65.79(11.59)	67.07 (10.93)	16710	0.565
Male Gender	0	Count (%)	5565 (99.8%)	11 (0.20%)	5576	0.418
	1	Count (%)	11118 (99.9%)	16 (0.14%)	11134	
BMI		Average (SD)	27.36 (4.04)	27.15 (3.44)		0.786
Cardiovascular Risk Factors						
Diabetes Mellitus	0	Count (%)	11988 (99.8%)	19 (0.16%)	12007	0.864
	1	Count (%)	4695 (99.8%)	8 (0.17%)	4703	
Hypertension	0	Count (%)	4493 (99.9%)	6 (0.1%)	4499	0.582
	1	Count (%)	12190 (99.8%)	21 (0.17%)	12211	
Smoking	0	Count (%)	10092 (99.8%)	17 (0.2%)	10109	0.793
	1	Count (%)	6591 (99.8%)	10 (0.15%)	6601	
Dyslipedemia	0	Count (%)	6665 (99.9%)	6 (0.09%)	6671	0.068
	1	Count (%)	10018 (99.8%)	21 (0.21%)	10039	
Prior Clinical History						
MI	0	Count (%)	13118 (99.8%)	23 (0.18%)	13141	0.410
	1	Count (%)	3565 (99.9%)	4 (0.11%)	3569	
PCI	0	Count (%)	12758 (99.8%)	22 (0.17%)	12780	0.541
	1	Count (%)	3925 (99.9%)	5 (0.13%)	3930	
CABG	0	Count (%)	15149 (99.8%)	24 (0.16%)	15173	0.731
	1	Count (%)	1534 (99.8%)	3 (0.20%)	1537	

Stroke/TIA	0	Count (%)	15399 (99.8%)	27 (0.18%)	15426	0.985
	1	Count (%)	1284 (100.0%)	0 (0.00%)	1284	
PAD	0	Count (%)	14765 (99.8%)	25 (0.17%)	14790	0.536
	1	Count (%)	1111 (99.9%)	1 (0.09%)	1112	
Non CABG Surgery	0	Count (%)	16259 (99.8%)	26 (0.16%)	16285	0.703
	1	Count (%)	424 (99.8%)	1 (0.24%)	425	
Renal Transplant	0	Count (%)	16613 (99.8%)	27 (0.2%)	16640	0.988
	1	Count (%)	70 (100.0%)	0 (0.00%)	70	
Clinical Setting and Procedural Characteristics						
Interventional procedure	0	Count (%)	9722 (99.8%)	15 (0.15%)	9737	0.775
	1	Count (%)	6961 (99.8%)	12 (0.17%)	6973	
Acute Coronary Syndrome	0	Count (%)	10485 (99.8%)	21 (0.20%)	10506	0.117
	1	Count (%)	6198 (99.9%)	6 (0.10%)	6204	
Aortic Valvulopathy	0	Count (%)	15465 (99.8%)	25 (0.16%)	15490	0.983
	1	Count (%)	1218 (99.8%)	2 (0.16%)	1220	
Systolic Blood Pressure	QC	Average (SD)	139.14 (28.16)	146.08 (29.26)	16489	0.209
Diastolic Blood Pressure	QC	Average (SD)	72.27 (13.05)	76.42 (15.16)	15902	0.104
Coronary artery disease extension	QD	min./median/max.	0/1/3	0/1/3	16710	0.932
Number of Treated Segments	QD	min./median/max.	0/1/3	0/1/3	15902	0.748
Fluoroscopy Time	QC	min./median/max.	0.24/7/105	1.45/6.40/40.30	16039	0.477
Contrast Volume	QC	min./median/max.	10/100/1100	40/91/300	15269	0.584
Number of Catheters Used	QD	min./median/max.	1/3/15	1/3/6	15902	0.972
Transradial Approach	0	Count (%)	12459 (99.8%)	20 (0.16%)	12479	
	1	Count (%)	4224 (99.8%)	7 (0.17%)	4231	0.942

QC: Quantitative Continuous; QD: Quantitative Discrete

Table 4.2: Odds ratio estimates and the corresponding 95% confidence intervals (CI) from the Univariate logistic regression, regarding the variables under study with a *p*-value less than 0.25 and the exposure (Transradial Approach).

Variable	OR	95 % CI	
		Lower	Higher
Dyslipidemia	2.33	0.93	5.77
Acute Coronary Syndrome	0.48	0.19	1.20
Systolic Blood Pressure	1.01	0.995	1.022
Diastolic Blood Pressure	1.02	0.995	1.053
Transradial Approach (Exposure)	1.03	0.44	2.44

The variables included in the preliminary model are: Dyslipidemia, the Acute Coronary Syndrome history, the Systolic and Diastolic Blood Pressure levels. The exposure variable Transradial Approach was included too. Although the variable does not have statistical significance, it is included in the model because it is our objective to access its contribution to the outcome. The OR and the respective 95% confidence interval, for every variable in the preliminary model are displayed in Table 4.2.

The univariate logistic regression only produces reliable results if there is a linear relation between the values of the variable and the outcome probability logit ($log\frac{p}{1-p}$). If this assumption does not hold, and one only relies on the logistic regression to access the importance of some variable, there is the risk to undervalue its contribution (Hastie and Tibshirani, 1986). And, as bad as undervalue its contribution,

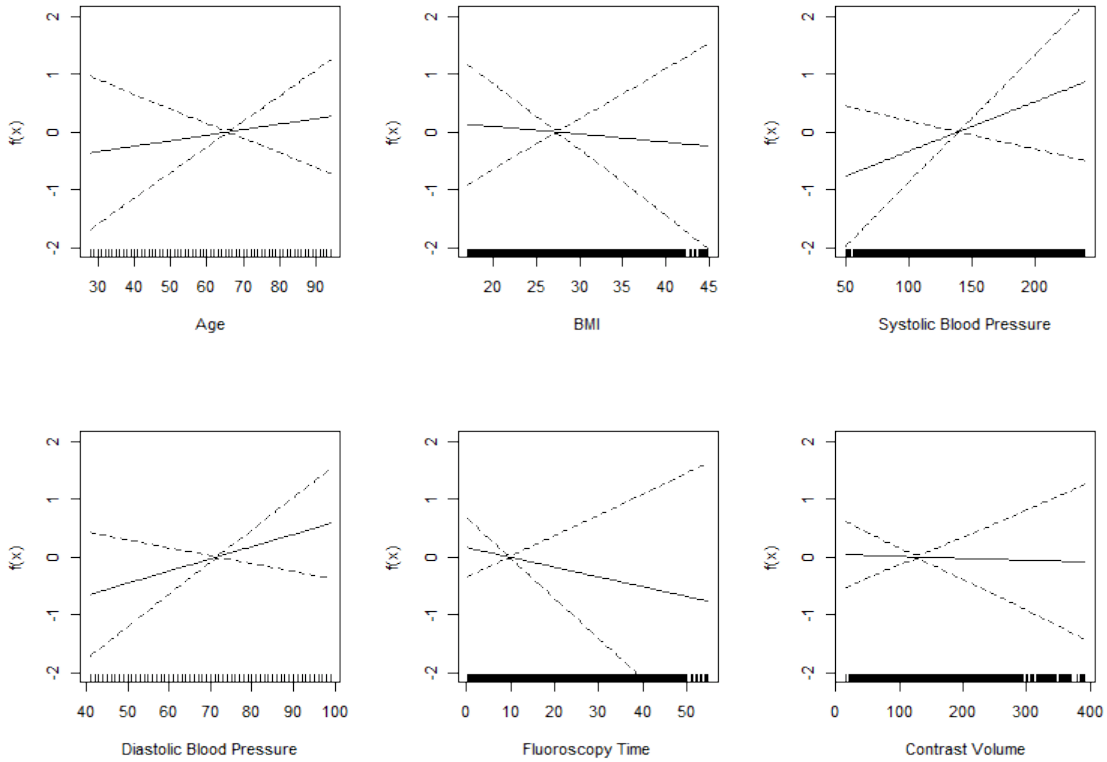


Figure 4.1: Univariate GAM functional forms regarding Age, BMI, Systolic BP, Diastolic BP, FLuoroscopy Time and Contrast Volume as covariates and Stroke/TIA as the dependent one. The functional form is represented by the filled lines and the confidence intervals are represented by the dashed lines.

is to build a model which do not represent the true relationship between the variable and the outcome (Hastie and Tibshirani, 1986). For this reason, an univariate GAM was performed to access the linearity assumption of each continuous variable. All of these variables proved to be linear relative to the logit of the outcome probability and, so, the results from the univariate logistic regression results are reliable (Figure 4.1). The splines used here and throughout this work are thin-plate regression splines.

If some continuous variable is not statistically significant in a LR context, it does not mean that it cannot provide relevant information. Because of this, categorizing variables can be useful to explain the outcome response. There are some variables, that otherwise would never be statistically significant if they remained continuous (Mazumdar and Glassman, 2000).

In an attempt to categorize some continuous variables and check for their significance, a minimum p-value approach was performed. The minimum p-values and their respective cutpoints are shown in Table 4.3, and the whole process is depicted in Figure 4.2. The variables Age, Systolic Blood Pressure, Diastolic Blood Pressure and Fluoroscopy Time have p-values (from the χ^2 tests) lower than 0.25 at their respective cutpoints. For this reason, the categorized versions of these variables (Age65, Systolic Blood Pressure147, Diastolic Blood Pressure72 and Fluoroscopy Time10) will be included in the preliminary model.

Of course, it makes no sense including Diastolic Blood Pressure and its categorized analog in the same model. So, it was decided to include the categorical variable Diastolic Blood Pressure72 because throughout the model selection process, the models with this variable had more discriminative power (measured by AUC), and roughly the same predictive power (measured by calibration plots) than the

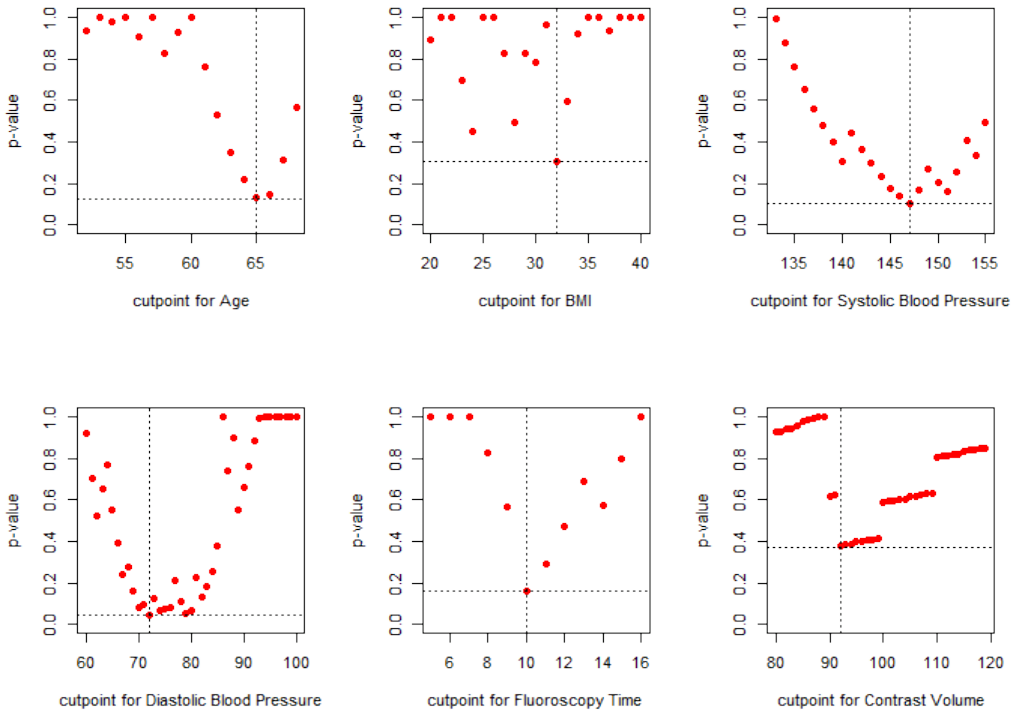


Figure 4.2: Minimum p-value approach to the quantitative continuous variables. Dashed lines represent the minimum p-value (horizontal), and the corresponding cutpoint (vertical).

model with its continuous counterpart. Regarding the Systolic Blood Pressure, the continuous analog was the chosen one for the same reason.

Table 4.3: Minimum p-value approach regarding Quantitative Continuous Variables

Variable	Cutpoint	p-value (χ^2 test)
Age	65	0.129
BMI	32	0.309
Systolic Blood Pressure	147	0.102
Diastolic Blood Pressure	72	0.046
Fluoroscopy Time	10	0.160
Contrast Volume	92	0.378

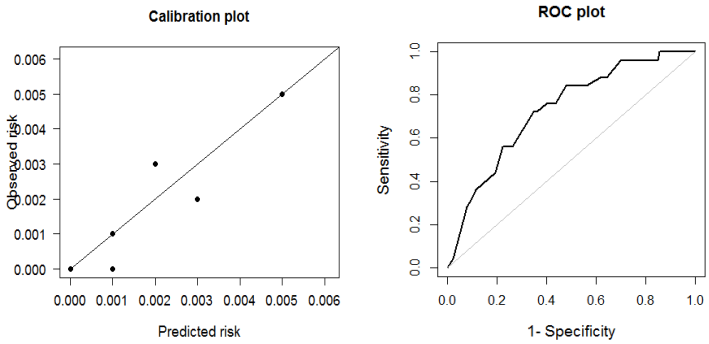
4.1.2 Multivariable Analysis

The preliminary model to describe the ocurrence of peri-procedural Stroke/TIA contains the following variables: Transradial Approach, Dyslipidemia, Acute Coronary Syndrome, Systolic Blood Pressure, Diastolic Blood Pressure72, Age65 and Fluoroscopy Time10. The final model was chosen, by removing the variables without statistical significance (p-value<0.10 for the Wald Test). All removing steps were based on decreasing AIC values (Backwards Stepwise). There were no identified interactions . The final model is depicted in Table 4.4 and is composed of the following variables: Dyslipidemia, Age65, Diastolic Blood Pressure72 and Fluoroscopy Time10 as the prognostically significant. The Hosmer-Lemeshow Test proved the model to be significant and well fitted because the Hosmer-Lemeshow Test p-value is 0.997

(Table 4.4). The calibration plot shows an acceptable predictive power (Figure 4.3a and Table 4.5) and the ROC curve with an AUC=0.731 shows a good discriminative power (Figure 4.3b and Table 4.4) .

Table 4.4: Final Model. Estimated coefficients, Odds Ratios, the corresponding 95% Confidence Intervals, the *p*-values from the Wald Tests, the Hosmer-Lemeshow *p*-value and the AUC value.

Variable	$\hat{\beta}$	$\hat{OR}$	95% CI		p-value	Hosmer	AUC
			Lower	Higher			
Transradial Approach	0.114	1.121	0.467	2.689	0.799		
Dyslipidemia	1.290	3.631	1.245	10.593	0.018		
Age65	0.810	2.247	0.959	5.265	0.062	0.997	0.731
Diastolic Blood Pressure72	0.976	2.653	1.133	6.208	0.025		
Fluoroscopy Time10	-0.885	0.413	0.154	1.102	0.077		



(a) Calibration plot (b) ROC curve

Figure 4.3: Calibration Plot and ROC curve regarding the model in Table 4.4.

Table 4.5: Hosmer-Lemeshow predicted risk bins. For each bin: Mean predicted and mean observed events, total predicted and total observed events. (model from Table 4.4).

Bins	Total	Mean Predicted	Mean Observed	Predicted	Observed
[0.000114,0.000301)	1660	0	0	0.4	0
[0.000301,0.000676)	2272	0	0	1.08	1
[0.000676,0.000729)	642	0.001	0	0.44	0
[0.000729,0.000998)	2107	0.001	0.001	1.73	3
0.000998	909	0.001	0	0.91	0
[0.001037,0.001834)	2123	0.001	0.001	2.68	3
[0.001834,0.002451)	1512	0.002	0.003	3.3	4
[0.002451,0.002643)	1006	0.002	0.003	2.49	3
[0.002643,0.002961)	1441	0.003	0.002	3.83	3
[0.002961,0.006629)	1505	0.005	0.005	8.13	8

4.1.3 Firth's Correction

In table 4.1, the Renal Transplant and the Stroke/TIA variables have zero values. Here we face a quasi-separation problem, in which the estimated coefficients have huge standard errors, and therefore causes these variables not to be statistically significant (p-values near 1). So, an univariate Firth's logistic regression was run, with Renal Transplant and Stroke/TIA as independent variables (Table 4.6). This quasi-separation problem did not arise when the categorization of Age, Diastolic Blood Pressure and Fluoroscopy Time was performed.

Table 4.6: Logistic Regression and Firth's Correction univariate analysis regarding β estimates, their standard errors (s.e.) and their p-values.

	Univariate Analysis						
Variable	Logistic Regression			Firth's Logistic Regression			
	$\hat{\beta}$	$\hat{\beta}$ s.e.	Wald Test p-value	$\hat{\beta}$	$\hat{\beta}$ s.e.	Wald's p-value	Likelihood Ratio p-value
Renal Transplant	-12.144	779.604	0.988	1.455	1.442	0.313	0.409
Stroke/TIA	-15.219	815.797	0.985	-1.523	1.428	0.286	0.162

The p-values decrease as expected in both variables. As the Stroke/TIA variable has a p-value lower than 0.25, this variable was included in a preliminary model (alongside with all variables with a p-value less than 0.25). The selection method was the same used in the previous logistic regression. In the end, Stroke/TIA is removed from the final model because is not statistically significant. The final model using Firth's correction contains the same variables as in the previous logistic regression, but with slightly different values regarding the β estimates and their standard errors (Table 4.7). The AUC is 0.731 (Table 4.7) and the calibration plot was very similar too (not shown), compared to the previous LR.

Table 4.7: Firth's Corrected Model. β coefficients, Odds Ratios, the corresponding 95% Confidence Intervals, p-values from the Likelihood Ratio test, p-value from the Hosmer-Lemeshow test and AUC.

Variable	$\hat{\beta}$	$\hat{OR}$	95% CI		Likelihood Ratio	Hosmer	AUC
			Lower	Higher	p-value		
Transradial Approach	0.156	1.168	0.468	2.633	0.723		
Dyslipidemia	1.195	3.304	1.307	10.512	0.010		
Age 65	0.778	2.177	0.973	5.254	0.058	0.934	0.731
Fluoroscopy Time10	-0.815	0.443	0.155	1.064	0.070		
Diastolic Blood Pressure72	0.944	2.570	1.151	6.196	0.021		

4.1.4 Treatment Effect Estimation (Odds Ratio)

The exposure to the Transradial or Transfemoral approach proved to be statistically non-significant (Table 4.4), and therefore does not contribute to the occurrence of peri-procedural Stroke/TIA. For anyone with Dyslipidemia, the odds are 263% higher to have a peri-procedural Stroke/TIA. Similarly, the odds are 124% higher for anyone with Age superior to 65 and 165% higher to anyone with Diastolic Blood Pressure superior to 72. The Fluoroscopy Time superior to 10 minutes is the only significant protector factor. If the Fluoroscopy Time (in minutes) is superior to 10, the odds are 58.7% lower than if the Fluorsocopy Time is lower than 10 minutes (Table 4.4).

4.2 Matching on the Propensity Score

Before matching, is clear that the treatment groups differ significantly regarding to some confounders and other variables (i.e. variables with SMD>0.10) (Table 4.8). This make it impossible to obtain an unbiased estimate comparing directly the outcomes of both groups. Our motivation is to create treatment groups that are as homogeneous as possible through matching on the Propensity Score (PS). The balance of variables that are suspected to influence the outcome (True Confounders and Potential Confounders) is essential (Austin, 2011a).

The potential confounding variables were the ones identified before as independent predictors of the outcome. The confounding variables were identified as the ones that are suspected to be associated with the outcome (in univariate LR, p-value<0.25), associated with the treatment assignment (in univariate LR, p-value<0.05), and the OR (regarding the exposure) is altered more than 10% when the variable is added to the model containing only the exposure as independent variable. The variables associated with the treatment assignment were identified as the ones that are independent predictors of the treatment assignment (p-value < 0.05) in a multivariable logistic regression context. The identified subsets of variables are displayed in Table 4.9.

Table 4.8: Standard Mean Difference (SMD) between Transradial and Transfemoral Group, before matching. Regarding four groups of variables: True Confounders, Potential Confounders, Variables Associated with Treatment Assignment and Other Variables.

Variable	Before Matching
True Confounders	
Systolic Blood Presure	0.324
Fluoroscopy Time	0.131
Potential Confounders	
Age	0.039
Dyslipidemia	0.008
Diastolic Blood Pressure	0.011
Fluoroscopy Time	0.131
Variables Associated with Treatment Assignment	
Age	0.039
BMI	0.078
Diabetes Mellitus	0.017
Dyslipidemia	0.008
MI	0.159
PCI	0.063
CABG	0.384
Stroke/TIA	0.071
Moderate/Severe CRD	0.089
Renal Transplant	0.085
Interventional Procedure	0.148
Acute Coronary Syndrome	0.024
Aortic Valvulopathy	0.156
Systolic Blood Presure	0.324
Diastolic Blood Pressure	0.011
Number of Treated Segments	0.204
Fluoroscopy Time	0.131
Coronary Artery Disease Extension	0.195
Contrast Volume	0.157
Other Variables	
Male Gender	0.017
Hypertension	0.03
Smoking	0.036
PAD	0.063
Non-CABG surgery	0.036
Number of Catheters Used	0.024

Table 4.9: Variable Subsets under study.

Subset	Variables
True Confounders	Systolic Blood Pressure, Fluoroscopy Time.
Potential Confounders	Age, Dyslipidemia, Systolic Blood Pressure, Fluoroscopy Time.
Variables Associated with Treatment Assignment	Age, BMI, Diabetes Mellitus, Dyslipidemia, MI, PCI, CABG, Stroke/TIA, Moderate/Severe CRD, Renal Transplant, Interventional Procedure, Acute Coronary Syndrome, Aortic Valvulopathy, Systolic Blood Pressure, Diastolic Blood Pressure, Coronary Artery Disease Extension, Number of Treated Segments, Fluoroscopy Time, Contrast Volume.

4.2.1 Propensity Score Model (Logistic Regression) and Matching

As stated in Section 3.5.3, there are four typical propensity score models. The variable subsets included in these models as independent covariates are: 1. True Confounders, 2. Potential Confounders, 3. Variables associated with the treatment assignment, 4. All observed variables. In this section, there were performed four matchings, using propensity scores estimated by logistic regression with the four possible variable subsets as independent covariates.

The categorized variables associated with the outcome and/or the treatment assignment were included in the PS models as continuous. This because balancing a continuous variable will balance its categorized analog, while the inverse is not true. Remind the balancing property $Z \perp \mathbf{X} | e(\mathbf{X})$ (Rubin and Rosenbaum, 1983).

In this context and throughout this work, the subjects subjected to Transradial Approach are the "Treated" Group, and the subjects subjected to Transfemoral Approach are the "Non-Treated" Group.

All matchings performed well, because all the variables targeted to be balanced in each matching were balanced (SMD<0.10) (Table 4.10), i.e. the matching performed with the PS estimated by the PS model with the True Confounders balanced the True Confounders between groups; the matching performed with the PS estimated by the PS model with the Potential Confounders balanced the Potential Confounders between groups and so on. This means that the PS estimates are consistent. Besides that, all the treated subjects were matched in all matching circumstances, which signals that the common region of support is not a problem. This was expected due to the great dimension of the Transfemoral Group compared to the Transradial Group.

The PS models that are most reassuring are the ones that contain All Variables and the Variables associated with Treatment Assignment, because these balanced all variables and not only the Confounders (taking into account the SMD Table 4.10). Analysing the QQ-plots (Figure 4.4), the matchings improved balance regarding Systolic BP, did not improve balance regarding Age and Diastolic BP (already balanced before matching), and made it worse regarding to Fluoroscopy Time.

Although they are both efficient, the PS model with only the Variables associated with Treatment Assignment, achieved a better balance regarding Potential and True Confounders which makes it a better

Table 4.10: Standard Mean Difference (SMD) between Transradial and Transfemoral Group, before matching and after matching. Propensity Scores used for matching were estimated by different propensity score models (True Confounders, Potential Confounders, Variables associated with treatment assginment and all observed variables). Matching with Replacement, recurring to NNM-CD with a 0.2 caliper.

Variable	Before Matching	After Matching			
			Variables within PS Model		
		All Variables	True Confounders	Potential Confounders	Variables Associated with Treatment Assignment
True Confounders					
Systolic Blood Presure	0.324	0.07	0.006	0.322	0.03
Fluoroscopy Time	0.131	0.016	0.014	0.017	0.002
Potential Confounders					
Age	0.039	0.02	0.118	0.063	0.002
Dyslipidemia	0.008	0.036	0.001	0.008	0.002
Diastolic Blood Pressure	0.011	0.024	0.105	0.028	0.033
Fluoroscopy Time	0.131	0.016	0.014	0.017	0.002
Variables Associated with Treatment Assignment					
Age	0.039	0.02	0.118	0.063	0.002
BMI	0.078	0.025	0.024	0.056	0.043
Diabetes Mellitus	0.017	0.008	0.01	0.052	0.011
Dyslipidemia	0.008	0.036	0.001	0.008	0.002
MI	0.159	0.024	0.174	0.17	0.015
PCI	0.063	0.012	0.108	0.058	0.026
CABG	0.384	0.006	0.422	0.399	0.025
Stroke/TIA	0.071	0.006	0.071	0.077	0.008
Moderate/Severe CRD	0.089	0.006	0.064	0.131	0.002
Renal Transplant	0.085	<0.001	0.067	0.056	0.032
Interventional Procedure	0.148	0.011	0.243	0.265	0.047
Acute Coronary Syndrome	0.024	0.033	0.06	0.034	0.023
Aortic Valvulopathy	0.156	0.035	0.177	0.14	0.001
Systolic Blood Presure	0.324	0.07	0.006	0.322	0.03
Diastolic Blood Pressure	0.011	0.016	0.105	0.028	0.033
Number of Treated Segments	0.204	0.01	0.254	0.266	0.023
Fluoroscopy Time	0.131	0.016	0.014	0.017	0.002
Coronary Artery Disease Extension	0.195	0.01	0.252	0.265	0.024
Contrast Volume	0.157	0.003	0.254	0.273	0.028
Other Variables					
Male Gender	0.017	0.003	0.088	0.013	0.003
Hypertension	0.03	0.001	0.001	0.084	0.001
Smoking	0.036	0.003	0.118	0.033	0.023
PAD	0.063	0.037	0.078	0.088	0.008
Non-CABG surgery	0.036	0.014	0.023	0.047	0.035
Number of Catheters Used	0.024	0.001	0.095	0.098	0.004

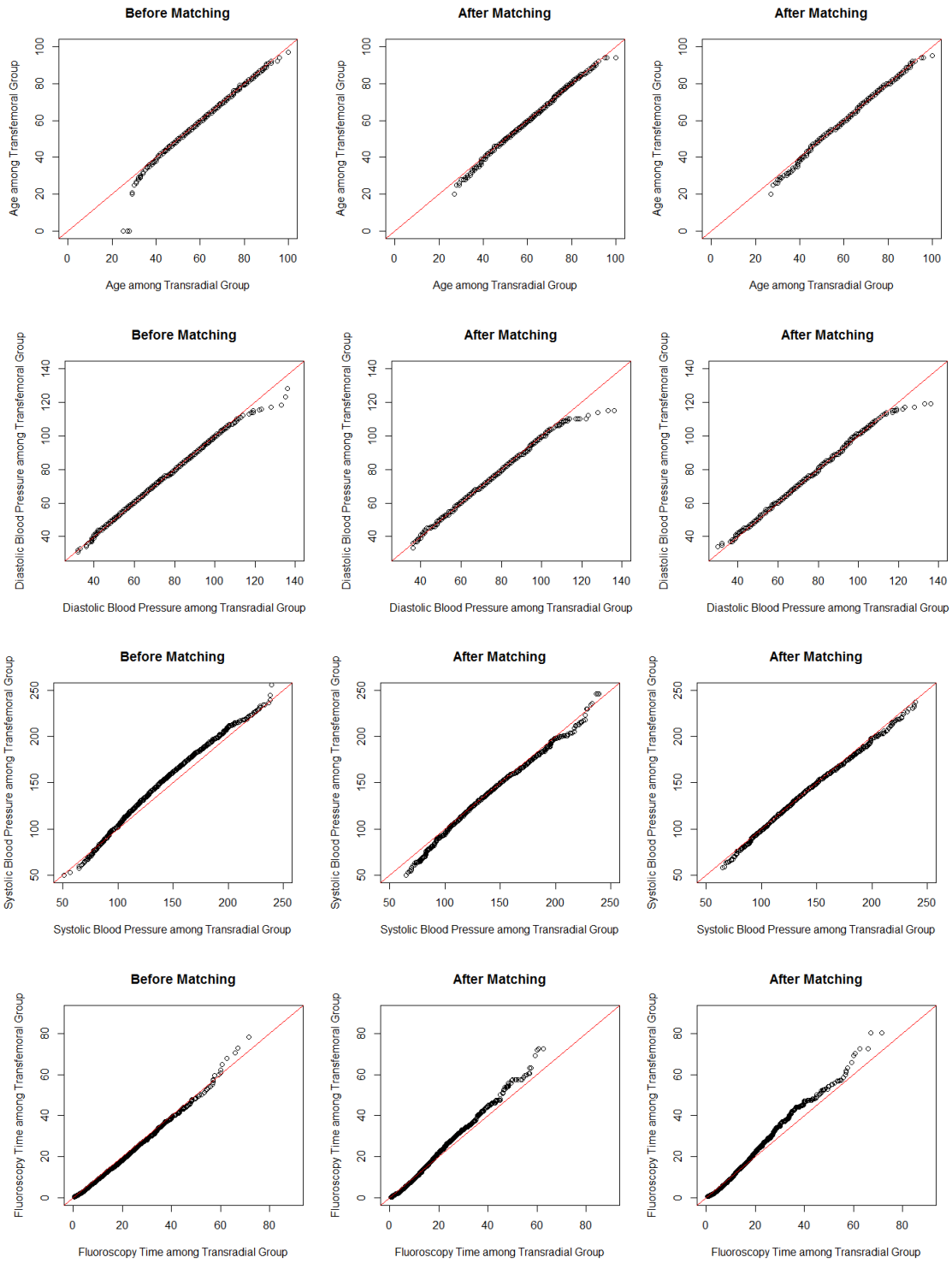


Figure 4.4: QQ-plots before and after matching regarding Age, Systolic Blood Pressure, Diastolic Blood Pressure and Fluoroscopy Time. (a) QQ-plots before matching. After matching, the LR PS model used to match is (b) the one containing all variables, (c) the one containing all the variables associated with treatment assignment.

choice for further analysis. This choice was based on a careful analysis of the QQ-plots and the SMD Table. The QQ-plots have more influence in this decision because they offer a more precise view of a variable distribution by treatment group.

4.2.2 Propensity Score Model and Matching: GAM Regression

Despite of the reasonable good balance achieved, there is still room for improvement. Applying GAMs to estimate PS may improve even more the balance because it provides better estimates due to its flexibility describing the true relationship between the independent covariates and the dependent one (Treatment Assignment). The motivation to use GAMs is based on the fact that Systolic Blood Pressure and Fluoroscopy Time are not linear with the treatment assignment logit (Figure 4.5), and therefore the LR PS estimates are not trustworthy. Because of that, it is natural that GAMs used to estimate the PS have, clearly, a superior predictive and discriminative power than LR models used before (Figure 4.6).

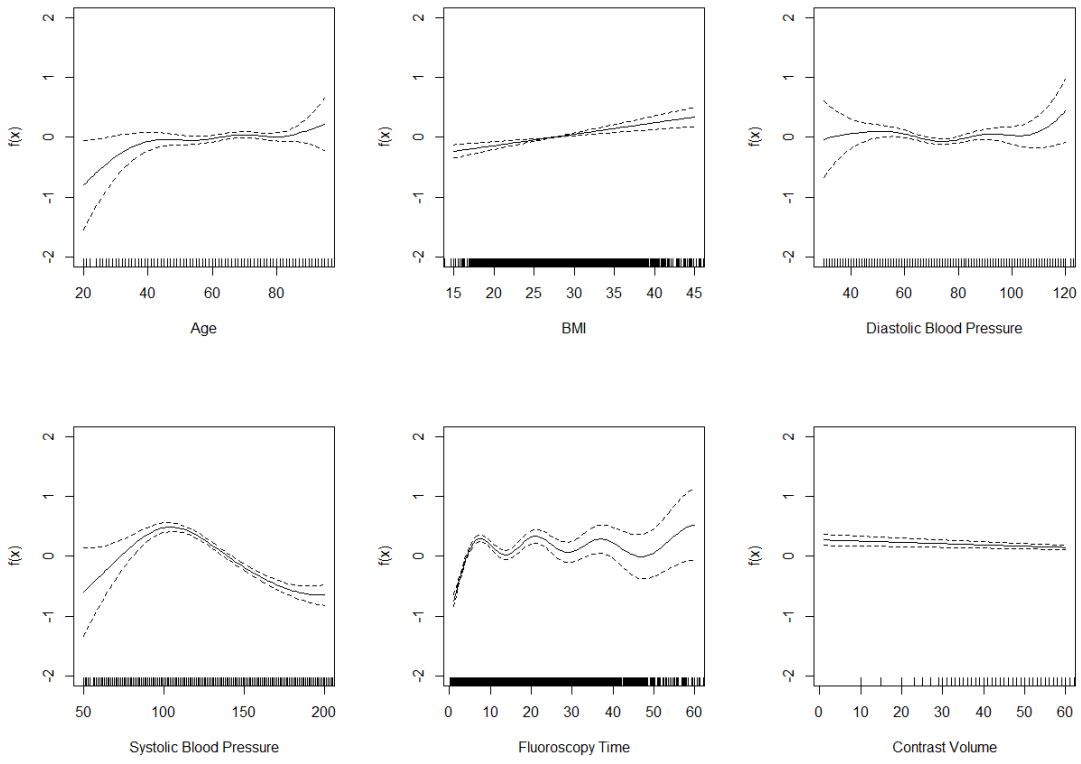


Figure 4.5: Univariate GAM functional forms regarding Age, BMI, Systolic BP, Diastolic BP, FLuoroscopy Time and Contrast Volume as covariates and Transradial Approach as the dependent one. The functional form is represented by the filled lines and the confidence intervals are represented by dashed lines.

New matching procedures were performed. Propensity scores were estimated based on two GAMs with the following subset of independent covariates:

- All observed variables
- Variables associated with Treatment Assignment

All the matchings on the GAMs PS estimates provided SMD<0.10, for all variables (Table 4.11). GAMs PS estimates provided more balanced matched samples, and therefore more successful matchings, when compared to the matchings on the LR PS estimates. This improvement is not observed while analysing the SMD Table (Table 4.11), but it is significant when looking to the QQ-plots (Figure 4.7).

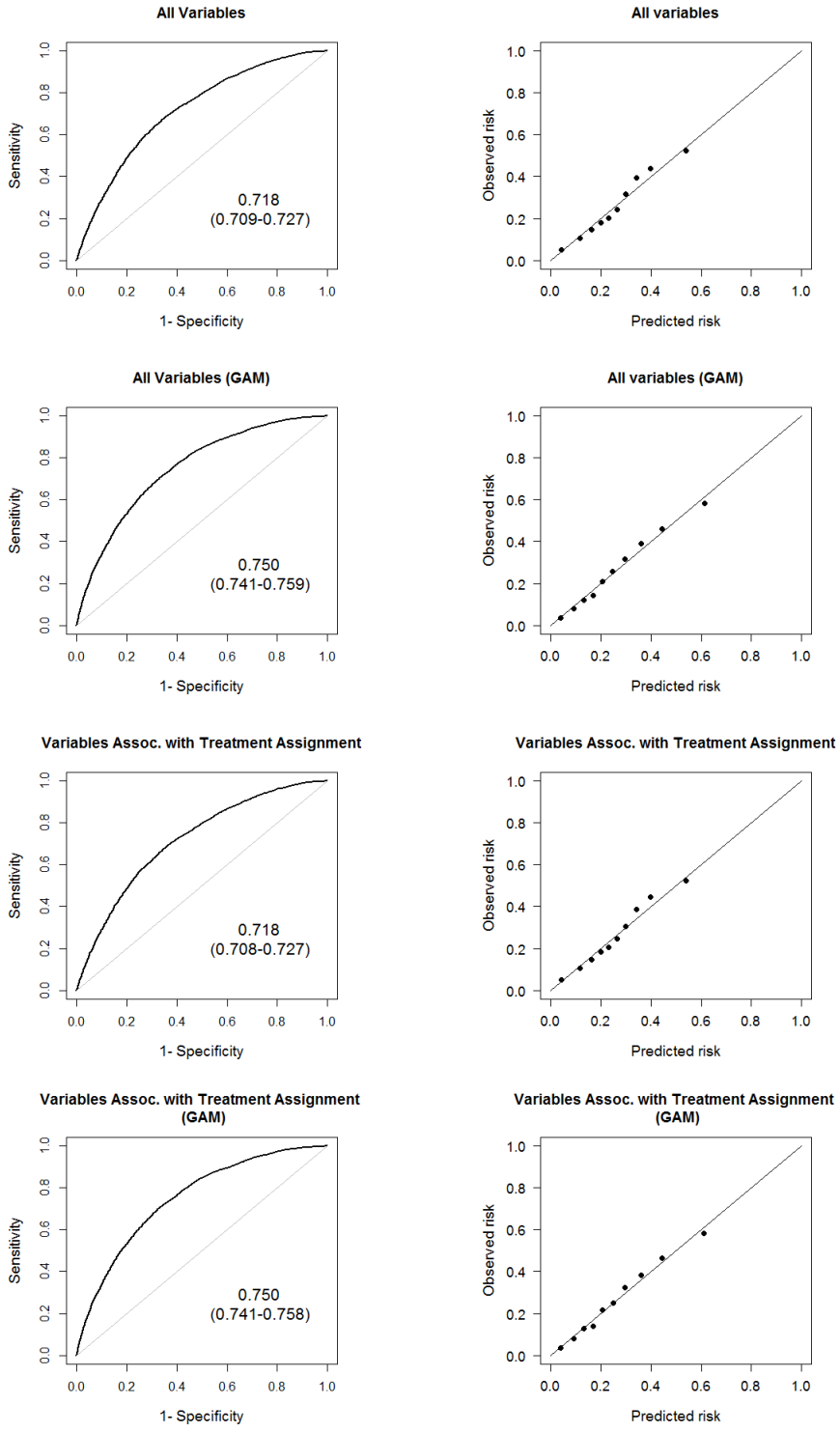


Figure 4.6: ROC plots and Calibration Plots regarding different PS models, with All Variables, All Variables (GAM), Variables Associated with Treatment Assignment, Variables Associated with Treatment Assignment (GAM). AUC and its (CI) are shown within each ROC plot.

Fluoroscopy Time and Systolic BP balance is very improved and Age balance is slightly improved, when matching on the GAMs estimates.

Between the GAMs, the PS estimates that provide the most balanced matched samples are the ones from the GAM with All Variables. Once more, this choice is based on a careful analysis on the SMD table

Table 4.11: Standard Mean Difference (SMD) between Transradial and Transfemoral Group, after matching. Propensity Scores used for matching estimated by different propensity score models (Variables associated with treatment assginment and All observed variables), with and without GAM regression. Matching with Replacement, recurring to NNM-CD with a 0.2 caliper.

Variable	**After Matching**			
		Variables within PS model		
	All Variables	All Variables (GAM)	Var. Associated with Treatment Assignment	Var. Associated with Treatment Assignment (GAM)
True Confounders				
Systolic Blood Presure	0.07	0.001	0.03	0.001
Fluoroscopy Time	0.016	0.022	0.002	0.084
Potential Confounders				
Age	0.02	0.010	0.002	0.029
Dyslipidemia	0.036	0.006	0.002	0.028
Diastolic Blood Pressure	0.024	0.015	0.033	0.03
Fluoroscopy Time	0.016	0.022	0.002	0.084
Variables Associated with Treatment Assignment				
Age	0.02	0.010	0.002	0.029
BMI	0.025	0.003	0.043	0.021
Diabetes Mellitus	0.008	0.001	0.011	0.043
Dyslipidemia	0.036	0.006	0.002	0.028
MI	0.024	0.010	0.015	0.021
PCI	0.012	0.022	0.026	0.006
CABG	0.006	0.034	0.025	0.023
Stroke/TIA	0.006	0.018	0.008	0.015
Moderate/Severe CRD	0.006	0.013	0.002	0.006
Renal Transplant	<0.001	0.026	0.032	0.019
Interventional Procedure	0.011	0.024	0.047	0.061
Acute Coronary Syndrome	0.033	0.016	0.023	0.016
Aortic Valvulopathy	0.035	0.017	0.001	0.008
Systolic Blood Presure	0.07	0.001	0.03	0.001
Diastolic Blood Pressure	0.016	0.015	0.033	0.03
Number of Treated Segments	0.01	0.013	0.023	0.016
Fluoroscopy Time	0.016	0.022	0.002	0.084
Coronary Artery Disease Extension	0.01	0.011	0.024	0.017
Contrast Volume	0.003	0.030	0.028	0.052
Other Variables				
Male Gender	0.003	0.006	0.003	0.019
Hypertension	0.001	0.016	0.001	0.013
Smoking	0.003	0.011	0.023	0.018
PAD	0.037	0.017	0.008	0.003
Non-CABG surgery	0.014	0.009	0.035	0.027
Number of Catheters Used	0.001	0.002	0.004	0.051

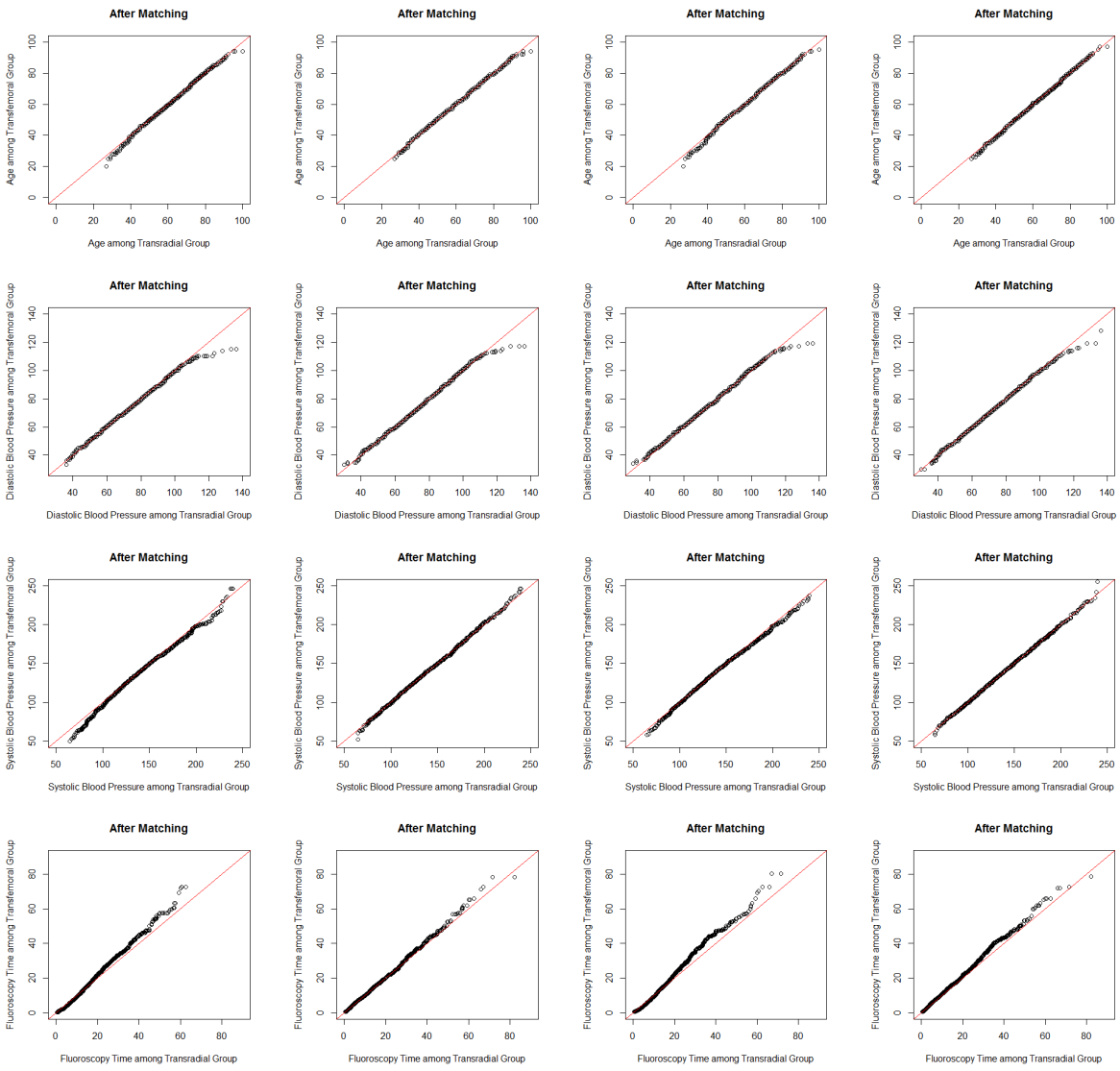


Figure 4.7: QQ-plots after matching regarding Age, Systolic Blood Pressure, Diastolic Blood Pressure and Fluoroscopy Time. The PS estimates used for matching are from: (a) LR containing all variables (b) GAM containing all variables, (c) LR containing the variables associated with treatment assignment, (d) GAM containing the variables associated with treatment assignment.

and the QQ-plots, because the two GAMs estimates perform in a very similar way. The main reason this model was chosen was the fact that its estimates made Fluoroscopy Time slightly more balanced than the other tested GAM with only the Variables Associated with Treatment Assignment.

4.2.3 Genetic Matching

Once again, in an attempt to improve balance, genetic matching was performed. The variable subsets used in the Generalized Mahalanobis Distance were:

- All observed variables + LR PS estimates (Estimated from the PS model with All observed variables as covariates) ;
- All observed variables + GAM PS estimates;
- All variables associated with treatment; assignment + LR PS estimates (Estimated from the PS model with All variables associated with treatment assignment as covariates) ;

- All variables associated with treatment assignment + GAM PS estimates.

As stated before, the matching is not done based on the crude predicted probability of assignment, but on the transformation log((1-x)/x). This transformation turns the crude estimate values distribution into a more normal one. This is of extreme importance when genetic matching is executed, because the MD perform better on normal distributed variables, that consequently will form a joint ellipsoidal distribution. This way, all non-ellipsoidal distributions can be matched successfully because they can be matched through the PS estimates. All variables present in this data set are non-normal except for Age, BMI, Systolic BP and Diastolic BP. (Figure 4.8).

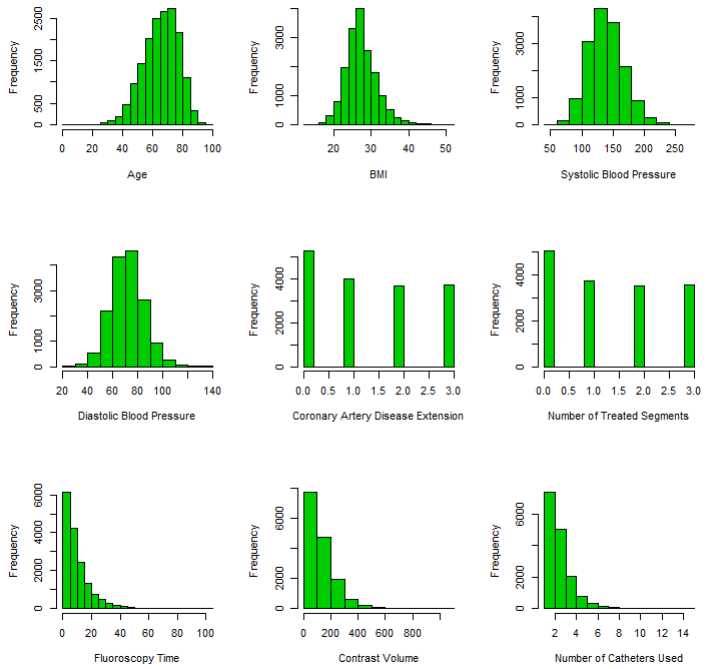


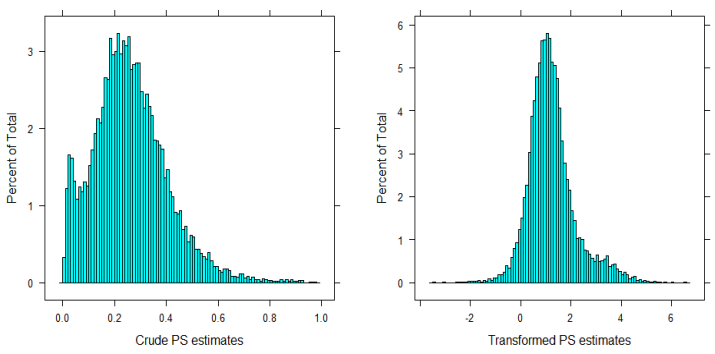
Figure 4.8: Histograms regarding Age, BMI, Systolic Blood Pressure, Diastolic Blood Pressure, Fluoroscopy Time, Contrast Volume and Number of Catheters Used. Bar Diagrams regarding Coronary Artery Decease and Number of Treated Segments,

All the PS transformations were very successful because they turned the crude PS estimates distributions from all PS models into normal distributed ones. (Results only shown for LR PS estimates with All Variables) (Figure 4.9).

In general, the genetic matchings were not significantly better than the matchings on the GAMs PS estimates. The GMD weights relative to the various genetic matchings are in Table 4.13.

All the genetic matchings provided SMD<0.10, for all variables (Table 4.12). The genetic matching using all variables plus the LR PS estimates and the genetic matching using the variables associated with treatment assignment plus the GAM PS estimates provided unbalanced matched samples regarding Systolic and Diastolic BP (Figure 4.10), so these matchings are discarded. The most successful genetic matching was the one with all variables plus the GAM PS estimates. This one achieved a higher balance on Age than any other.

The weights regarding each variable in each genetic matching are shown in Table 4.13.

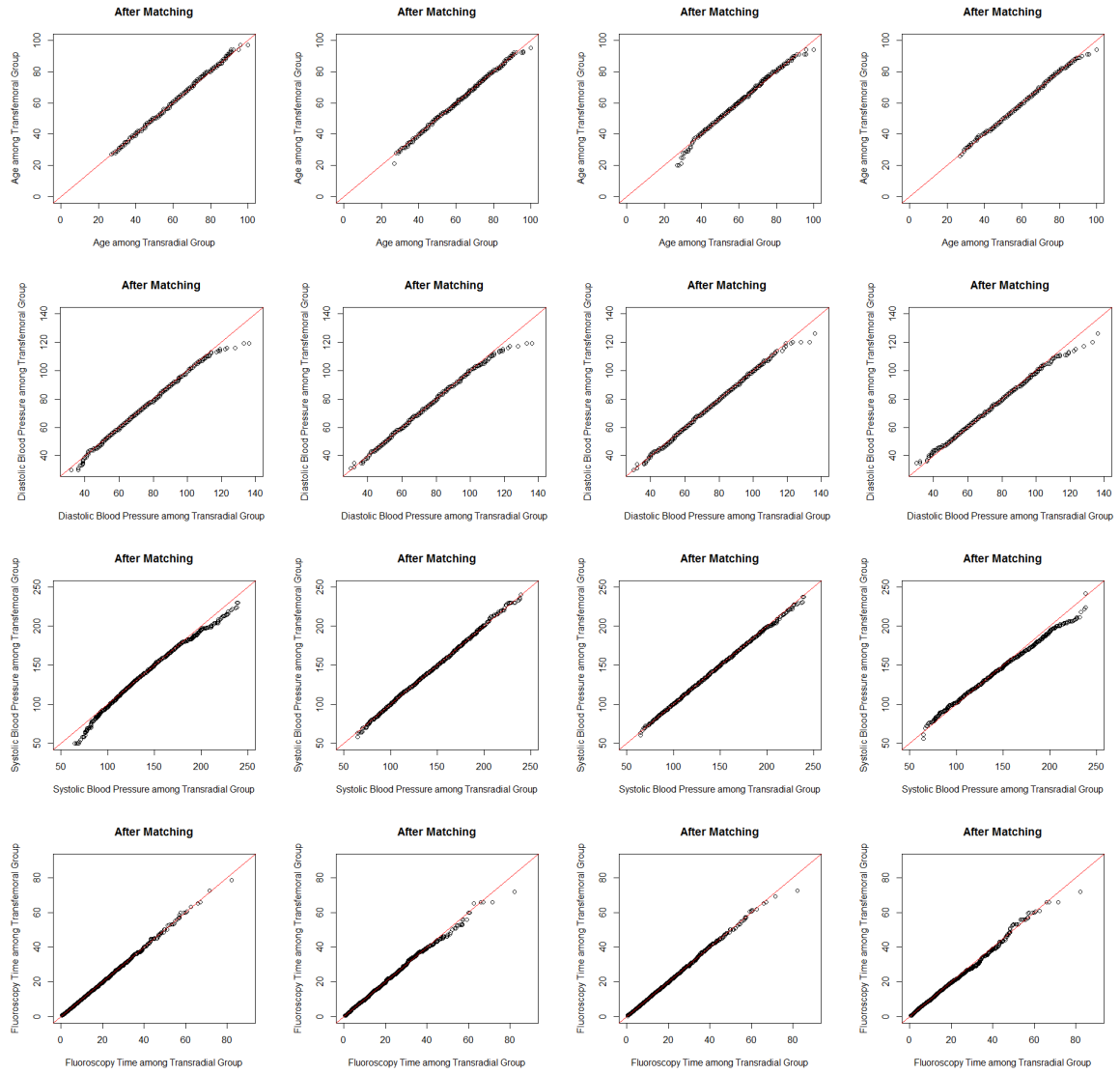


(a) (b)

Figure 4.9: Histograms of the: (a) Crude PS estimates and the (b) Transformed PS estimates. Estimated from the LR PS model with All Variables.

Table 4.12: Standard Mean Difference (SMD) between Transradial and Transfemoral Group, after Genetic matching (pop. size=300). Propensity Scores estimated by different propensity score models with Variables associated with treatment assginment and All observed variables, modeled by LR and GAM. Matching with Replacement, recurring to NNM-CD with a 0.2 caliper. Loss Function: Maximum Kolmogorov-Smirnov statistic.

Variable	**After Genetic Matching**			
	Variable Subsets + PS estimates			
	All Variables + LR PS estimates	All Variables + GAM PS estimates	Var. Associated with Treatment Assignment + LR PS estimates	Var. Associated with Treatment Assignment + GAM PS estimates
True Confounders				
Systolic Blood Presure	0.039	0.004	0.007	0.011
Fluoroscopy Time	0.011	0.005	0.009	0.021
Potential Confounders				
Age	0.05	0.002	0.054	0.01
Dyslipidemia	0.044	0.022	0.001	0.017
Diastolic Blood Pressure	0.046	0.01	0.038	0.012
Fluoroscopy Time	0.011	0.005	0.009	0.021
Variables Associated with Treatment Assignment				
Age	0.05	0.002	0.054	0.01
BMI	0.054	0.017	0.031	0.002
Diabetes Mellitus	0.002	0.014	0.001	0.024
Dyslipidemia	0.044	0.022	0.001	0.017
MI	0.013	<0.001	0.008	0.009
PCI	0.01	0.005	0.028	0.001
CABG	0.004	0.004	0.015	0.011
Stroke/TIA	0.02	0.045	0.043	0.017
Moderate/Severe CRD	0.022	0.018	<0.001	0.011
Renal Transplant	0.013	<0.001	0.026	0.019
Interventional Procedure	0.068	0.016	0.043	0.018
Acute Coronary Syndrome	0.057	0.021	0.04	<0.001
Aortic Valvulopathy	0.006	0.004	<0.001	0.007
Systolic Blood Presure	0.039	0.004	0.007	0.011
Diastolic Blood Pressure	0.046	0.01	0.038	0.012
Number of Treated Segments	0.001	0.016	0.035	0.016
Fluoroscopy Time	0.011	0.005	0.009	0.021
Coronary Artery Disease Extension	0.001	0.017	0.036	0.015
Contrast Volume	0.011	0.011	0.015	0.009
Other Variables				
Male Gender	0.01	0.003	0.012	0.027
Hypertension	0.011	0.008	0.003	0.005
Smoking	0.037	0.004	0.001	0.011
PAD	0.002	0.009	0.017	0.027
Non-CABG surgery	0.013	0.031	0.004	0.028
Number of Catheters Used	0.004	<0.001	0.046	0.001



(a) (b) (c) (d)

Figure 4.10: QQ-plots after matching regarding Age, Systolic Blood Pressure, Diastolic Blood Pressure and Fluoroscopy Time. The matching method is: (a) genetic containing all variables plus the LR PS estimates, (b) genetic containing all variables plus the GAM PS estimates, (c) genetic containing variables associated with treatment assignment plus the LR PS estimates, (d) genetic containing variables associated with treatment assignment plus the GAM PS estimates.

Table 4.13: Weights regarding Generalized Mahalanobis Distance (GMD), for every Genetic Matching performed [Variable Subset + PS estimates]. SMD for every variable before matching.

Variable	All Variables + LR PS estimates	All Variables + GAM PS estimates	Var. Associated with Treatment Assignment + LR PS estimates	Var. Associated with Treatment Assignment + GAM PS estimates	SMD Before Matching
Age	0.0	0.0	0.0	0.0	0.039
Male Gender	0.0	0.0			0.017
BMI	0.0	0.0	0.0	132.5	0.078
Diabetes Mellitus	0.0	0.0	0.0	0.0	0.017
Hypertension	0.0	0.0			0.03
Smoking	2.1	0.0			0.036
Dyslipidemia	0.0	0.0	0.0	0.0	0.008
MI	0.0	981.5	0.0	0.0	0.159
PCI	0.0	0.0	0.1	0.0	0.063
CABG	33.9	0.0	0.0	0.0	0.384
Stroke TIA prev	0.0	0.0	0.0	0.0	0.071
PAD	0.0	0.0			0.063
Non CABG Surgery	0.0	0.0			0.036
Moderate Severe CRD	0.0	0.0	208.2	0.0	0.089
Renal Transplant	0.0	572.2	0.0	0.0	0.085
Interventional Procedure	0.0	0.0	0.0	0.0	0.148
Acute Coronary Syndrome	0.0	0.0	0.0	428.7	0.024
Aortic Valvulopathy	0.0	0.0	110.2	0.0	0.156
Systolic Blood Pressure	0.0	0.0	142.9	2.9	0.324
Diastolic Blood Pressure	0.0	0.0	0.0	0.0	0.011
Coronary Artery Disease Ext.	22.5	0.0			0.195
Number of Treated Segments	785.7	0.0	0.0	0.0	0.204
Fluoroscopy Time	829.8	0.0	106.8	0.0	0.131
Contrast Volume	175.9	0.0	0.0	0.0	0.157
Number of Catheters Used	724.5	491.9	0.0	0.0	0.024
Propensity Score	848.3	797.9	907.0	995.1	

4.2.4 Treatment Effect Estimation (Average Treatment Effect for the Treated)

Finally, there is the need to choose the best matching, in order to estimate the Average Treatment Effect for the Treated (ATT). Comparing the SMD tables (Tables 4.10, 4.11, 4.12) and the QQ-plots (Figures 4.4, 4.7, 4.10), the matching that achieved higher balance on the confounders was the Genetic Matching with all variables plus the GAM PS estimates. So, this will be the one used to estimate the ATT.

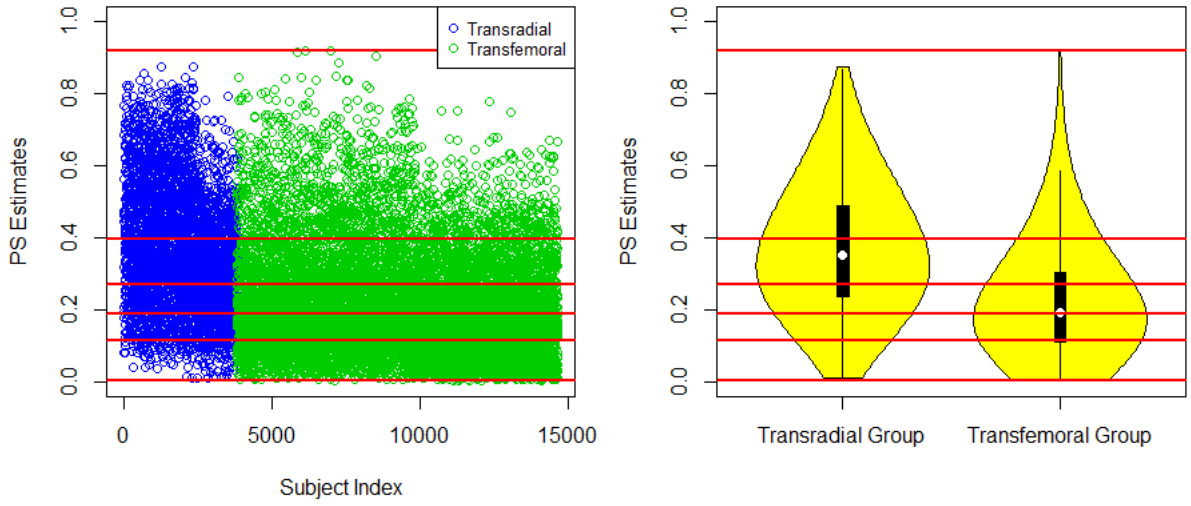
Table 4.14: Average Treatment Effect for the Treated and the respective Standard-Deviation (SD) estimated by Matching.

$\hat{\tau}_{ATT}$	$SD(\hat{\tau}_{ATT})$	p-value $H_0 : \tau_{ATT} = 0$
0.000786	0.00087	0.36566

The Matching ATT estimate is 0.000786. This is not a significative difference because the p-value is 0.36566 (Table 4.14).

4.3 Stratification

The PS estimates used to stratify are fitted values of a GAM with all observed variables as dependent covariates (Figure 4.11). Other PS models were tested like logistic regression with all variables, logistic regression with the variables associated with exposure and a GAM with the variables associated with exposure. The chosen PS estimates provide the best variable balance across all strata. Although there are some SMD's above 0.10, there are none above 0.25 (Table 4.15) which is the maximum acceptable value (Guo and Fraser, 2009).



(a) (b)

Figure 4.11: (a) PS estimates by Stratum, and by treatment group (Transradial coloured Blue and Transfemoral coloured Green). Stratum delimiters coloured Red. (b) Vioplots of the PS estimates of the two treatment groups. Vioplots are box-plot analogues which depict, additionally, the density function (coloured Yellow).

In Stratum 1, there are 3 confounders unbalanced (Fluoroscopy Time, Age, Dyslipidemia). In Stratum 2, there are 2 confounders unbalanced (Fluoroscopy Time, Diastolic Blood Pressure, Dyslipidemia). In Stratum 3 and 4, there are no confounders unbalanced. And in Stratum 5, 1 confounder unbalanced (Dyslipidemia).

Besides the confounders, in Stratum 1, there are more 4 unbalanced variables (CABG, Moderate/Severe CRD, PAD, Non-CABG surgery). In Stratum 2, more 2 unbalanced (Acute Coronary Syndrome, Aortic Valvulopathy). In Stratum 3 and 4, no more unbalanced variables. In Stratum 5, 1 unbalanced variable (Interventional Procedure).

A more efficient way to analyse the variable balance is to check for the QQ-plots. In this case, the continuous confounding variables (Age, Diastolic BP, Systolic BP and Fluoroscopy Time) were analysed through this method (Figure 4.12), all these variables seem to be reasonably balanced across all strata except for the Systolic BP. This do not contradict the low SMD regarding this variable across all strata (Table 4.15), because two different distributions may have equal means.

Table 4.15: Standard Mean Difference (SMD) between Transradial and Transfemoral Group, in each stratum. Propensity Scores used estimated by a GAM with All observed variables as dependent covariates.

Variable	**Stratification** (Interquintile Range)				
	Stratum1 $(0-1^{st})$	Stratum2 $(1^{st}-2^{nd})$	Stratum3 $(2^{nd}-3^{rd})$	Stratum4 $(3^{rd}-4^{th})$	Stratum5 $(4^{th}-5^{th})$
True Confounders					
Systolic Blood Presure	0.066	0.012	0.073	0.02	0.094
Fluoroscopy Time	0.107	0.104	0.06	0.024	0.081
Potential Confounders					
Age	0.106	0.056	0.016	0.093	0.042
Dyslipidemia	0.19	0.075	0.018	0.059	0.13
Diastolic Blood Pressure	0.08	0.148	0.055	0.061	0.008
Fluoroscopy Time	0.107	0.104	0.06	0.024	0.081
Variables Associated with Treatment Assignment					
Age	0.106	0.056	0.016	0.093	0.042
BMI	0.014	0.047	0.038	0.004	0.064
Diabetes Mellitus	0.034	0.006	0.022	0.006	¡0.001
Dyslipidemia	0.19	0.075	0.018	0.059	0.13
MI	0.067	0.034	0.026	0.052	0.028
PCI	<0.001	0.021	0.014	0.006	0.018
CABG	0.211	0.051	0.027	0.029	0.036
Stroke/TIA	0.033	0.034	0.021	0.002	0.018
Moderate/Severe CRD	0.149	0.076	0.019	0.027	0.029
Renal Transplant	0.069	0.028	$***$	0.033	$***$
Interventional Procedure	0.054	0.051	0.049	0.036	0.17
Acute Coronary Syndrome	0.055	0.151	0.01	0.092	0.076
Aortic Valvulopathy	0.012	0.134	0.017	0.026	0.053
Systolic Blood Presure	0.066	0.012	0.073	0.02	0.094
Diastolic Blood Pressure	0.08	0.148	0.055	0.061	0.008
Number of Treated Segments	0.07	0.08	0.009	0.06	0.066
Fluoroscopy Time	0.107	0.104	0.06	0.024	0.081
Coronary Artery Disease Extension	0.072	0.079	0.012	0.06	0.069
Contrast Volume	0.073	0.028	0.029	0.091	0.094
Other Variables					
Male Gender	0.087	0.018	0.041	0.043	0.099
Hypertension	0.001	0.034	0.002	0.028	0.037
Smoking	0.055	0.031	0.015	0.04	0.099
PAD	0.139	0.034	0.071	0.086	0.067
Non-CABG surgery	0.128	0.006	0.03	0.028	0.009
Number of Catheters Used	0.096	0.096	0.061	0.042	0.074

$***$ Renal Transplant did not have subjects in both treatment groups, in stratum 3 and 5.

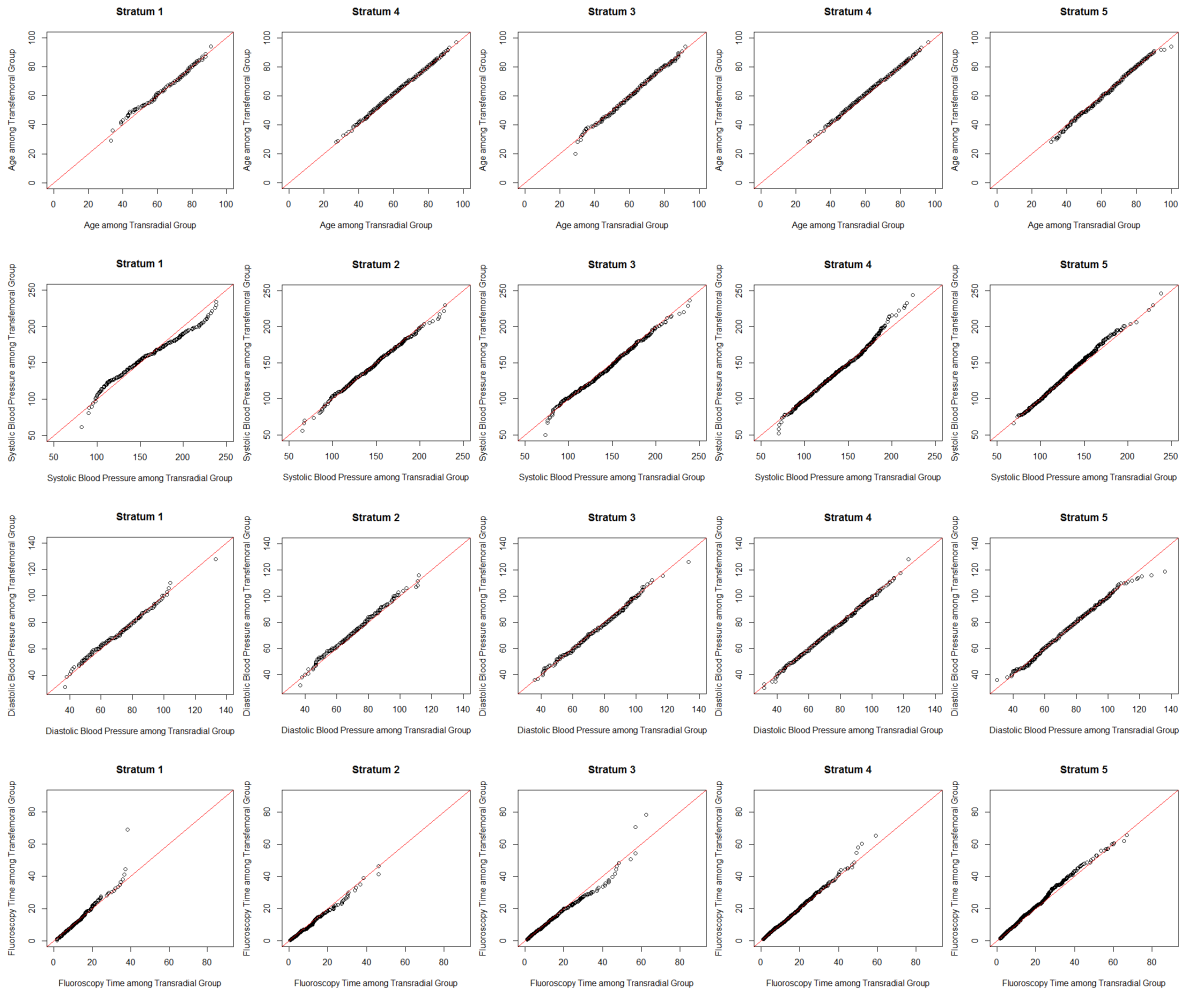


Figure 4.12: QQ-plots after stratification regarding Age, Systolic Blood Pressure, Diastolic Blood Pressure and Fluoroscopy Time, at each stratum. Stratum 1: Subjects within 0-1^{st} quintile, Stratum 2: Subjects within 1^{st}-2^{nd} quintile, Stratum 3: Subjects within 2^{nd}-3^{rd} quintile, Stratum 4: Subjects within 3^{rd}-4^{th} quintile, Stratum 5: Subjects within 4^{th}-5^{th} quintile.

4.3.1 Treatment Effect Estimation (Average Treatment Effect)

The Stratification Overall ATE estimate is -0.000386. This is not significant because the p-value is 0.50691 (Table 4.16).

Table 4.16: Average Treatment Effect estimates (Stratum-specific and Overall) and its Standard-Deviation (SD), estimated by Stratification .

	Stratum-Specific		Overall	Overall	p-value
	$\hat{\tau}_{ATE}$	$SD(\hat{\tau}_{ATE})$	$\hat{\tau}_{ATE}$	$SD(\hat{\tau}_{ATE})$	$H_0 : \tau_{ATE} = 0$
Stratum1	-0.00217	0.00088			
Stratum2	-0.00157	0.00078			
Stratum3	-0.00222	0.00099	-0.000386	0.00058	0.50692
Stratum4	0.00280	0.00206			
Stratum5	0.00125	0.00133			

4.4 Inverse Probability Treatment Weighting

The PS estimates used in IPTW were the fitted values of the logistic regression with Variables Associated with Treatment Assignment as covariates. This model was used because it provided sightly better variable balances than the logistic regression with All Variables, in PS matching (Chapter 4.2). These great balance achieved in Matching shows that these PS estimates are credible.

PS estimated by GAMs cannot be used in IPTW because there is no way to find the final $\hat{\tau}_{ATE}$ variance and so, significance tests are not possible (Lunceford and Davidian, 2004).

4.4.1 Treatment Effect Estimation (ATE)

Table 4.17: Average Treatment Effect and its Standard-Deviation (SD) estimated by IPTW .

$\hat{\tau}_{ATE}$	$SD(\hat{\tau}_{ATE})$	p-value $H_0 : \tau_{ATE} = 0$
-0.000155	0.00029	0.59150

The IPTW ATE estimate is -0.000155. This is not significant because the p-value is 0.59150. (Table 4.17)

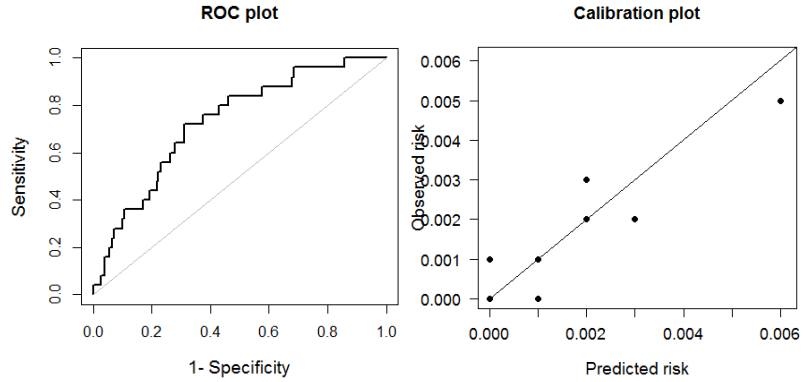
4.5 GLM and PS mixed approach: Covariate Adjustment

In an attempt to include all variables in the model, it was added to the final LR model obtained in 4.1.2, a ”summary” variable. This variable is the PS estimate obtained from a logistic regression PS model containing all the variables except for the ones already in the model (Dyslipidemia, Age, Fluoroscopy Time and Diastolic Blood Pressure). The model is described in Table 4.18; and the ROC Curve and Calibration Plot are displayed in Figure 4.13.

The PS estimate addition into the model does not provide any improvement, because the model’s discriminative and predictive power remain roughly the same. This suggests that the variables included through the PS are not valuable to explain the event occurrence. PS estimates were also obtained from a GAM and introduced in the model, which returned very similar results.

Table 4.18: LR Model with PS estimates as variable. β coefficients, Odds Ratios, its 95% confidence intervals, and the Wald Test p-values.

Variable	β	OR	95% CI		Wald p-value	Hosmer	AUC
			Lower	Higher			
Transradial Approach	0.1243	1.132	0.454	2.825	0.790	0.997	0.731
Dyslipidemia	1.284	3.609	1.235	10.550	0.019		
Age65	0.805	2.237	0.947	5.284	0.066		
Fluoroscopy Time10	-0.886	0.412	0.151	1.126	0.085		
Diastolic Blood Pressure72	0.965	2.625	1.098	6.273	0.030		
100×PS Estimates	-0.0013	0.999	0.957	1.042	0.954		



(a) (b)

Figure 4.13: (a) ROC curve regarding model in Table. b) Calibration Plot regarding the Hosmer-Lemeshow Test applied to model in Table 4.18

Table 4.19: Hosmer-Lemeshow predicted risk bins. For each bin: Mean predicted and mean observed events, total predicted and total observed events. Regarding model in Table 4.18.

Bins	Total	Mean Predicted	Mean Observed	Predicted	Observed
[0.000115,0.000298)	1466	0	0	0.36	0
[0.000298,0.000631)	1465	0	0.001	0.58	1
[0.000631,0.000737)	1466	0.001	0	0.98	0
[0.000737,0.000956)	1465	0.001	0.001	1.19	2
[0.000956,0.001062)	1466	0.001	0.001	1.48	1
[0.001062,0.001661)	1465	0.001	0.001	1.7	2
[0.001661,0.002316)	1466	0.002	0.002	2.94	3
[0.002316,0.002625)	1465	0.002	0.003	3.63	5
[0.002625,0.002990)	1466	0.003	0.002	4.01	3
[0.002990,0.007056]	1465	0.006	0.005	8.14	8

The Code used to produce the results in Chapter 4 is in the Appendices.

Chapter 5

Discussion

It was possible to identify some variables associated with the occurrence of a peri-procedural Stroke/TIA, within logistic regression. They were Age above 65 y, Dyslipidemia, higher Diastolic Blood Pressure (>72 mmHg) and Fluoroscopy Time. All of the identified variables make sense being included in the model. Relative do their younger counterparts, older people have a higher risk of all procedure-related complications because of comorbidity, vascular stiffness and tortuosity and a high burden of vascular disease. Dyslipidemia and high blood pressure, with no surprise, are important risk factors, also due to their association with more advanced atherosclerotic disease (Donnan et al., 2008; Sacco et al., 1997). Paradoxically, Fluoroscopy Time was the only identified protector factor. Fluoroscopy time is a surrogate of higher procedure duration and vascular manipulation, and thus it is inverse association with neurologic complications is apparently paradoxical. The reasons for this are not clear. This can be a *proxy* variable - meaning that it is not in itself directly relevant, but can be related to unobservable or immeasurable confounders - or simply a matter of statistical chance. It is arguable than, weather or not this variable should have been kept in the final model.

There are some variables that would make sense too be on the final model as Hypertension, Interventional Procedure, Aortic Valvulopathy, Acute Coronary Syndrome, Coronary Artery Disease Extension, Number of Treated Segments, Contrast Volume, Diabetes mellitus and Smoking History. Mostly, these variables are not in the model, because they are correlated with the ones included in it. High Systolic BP is, almost, Hypertension synonym. The variables Interventional Procedure, Aortic Valvulopathy, Acute Coronary Syndrome, Coronary Artery Disease Extension and Number of Treated Segments are all indicators of the general health condition of the patient, which is associated with the Age, Dyslipidemia and BP. The Smoking history and Diabetes mellitus are direct risk factors for the occurrence of a Stroke/TIA (Donnan et al., 2008; Sacco et al., 1997).

One of the main drawbacks of these study were the low number of events (27), which jeopardized and limited the quality of the model (Peduzzi et al., 1996). The validity of the logistic regression was verified by checking the linearity of the variables with the outcome logit, and by so a GAM was not required to model the data efficiently. There were some variables like Age and Fluoroscopy Time that were not significant as continuous variables but proved their value to the model by being categorized. This means that only major changes in these variables are worth being seen as risk factors (Mazumdar and Glassman, 2000).

As stated before, the low number of events harms the quality of the model because there is not enough information being considered. One of the problems that comes with this is the Separation

Problem (Heinze and Schemper, 2002). The variables affected by this were the Renal Transplant and the Stroke/TIA history. In fact, the Stroke/TIA history can be considered as a variable that can explain the outcome variable (Sacco et al., 1997) and, if not included in the final model because of the Separation problem, the final model is in risk of being misspecified. Although, it was not the case because the Firth's correction returned roughly the same model as the traditional logistic regression discussed in the paragraphs above.

Returning to the main objective of this work, the logistic regression results pointed that Transradial Approach is not a risk factor related to the ocurrence of peri-operational Stroke/TIA. This result is in accordance with the literature (Ratib et al., 2013a; Raposo et al., 2015). It is noteworthy that only the logistic regression enables us to separately identify different risk factors, contrary to the PS methodologies (Austin, 2011a).

Following the traditional Logistic Regression, this work moved into the estimation of new kinds of treatment effects besides the OR. One of them is the ATT, which is estimated based on the PS matching of subjects. The best matching is the one which balances the most prognostically important variables and if possible all the other variables, between treatment groups. Because of that, there is the need to search for the PS estimates that satisfy us the most (Guo and Fraser, 2009).

There were obtained very diverse PS estimates from various PS models (Logistic Regression and Generalized Additive Models), with different subsets of variables as independent covariates (True Confounders, Potential Confounders, Variables Associated with Treatment Assignment and All). The PS estimates that originated the best balance were the ones from the GAM PS model with all variables. This is not surprising because the Fluoroscopy Time and Systolic Blood Pressure are not linear with the treatment assignment logit, and so GAMs are a better fitting strategy (more predictive and discriminative power obtained) (Woo et al., 2008). That is why, after matching with GAM PS estimates, Fluoroscopy Time and Systolic Blood Pressure were the variables that improved more its balance between treatment groups. This GAM PS estimates summarize these variables in a more efficient way, and so, the balance regarding these variables improves.

The best subsets of variables to be included in the PS model were All and the ones Associated with Treatment Assignment. This is expected because these subsets balance all variables and not only the confounders. The subset with the Variables Associated with Treatment Assignment balances all variables because the ones that are unbalanced before matching are present in this subset, in the end, all observed variables stay balanced after matching.

In the search for a even better matching, Genetic Matching appears as a solid alternative (Diamond and Sekhon, 2012). Generally, it produced slightly more successful matchings. This is expected because it only matches on the variables if there is some advantage in doing so. If not, it continues to match on the PS estimates only. The best genetic matching was the one with All variables plus the GAM PS estimates (estimated from All variables as independent covariates). This is not a big surprise because GAM PS estimates with all the variables had already been considered to provide the best matching among the non-genetic matchings.

With no surprise the weights attributed to the PS estimates in all genetic matchings performed were very high, which proves their value on matching (Diamond and Sekhon, 2012). There are few categorical variables with weights different from zero, which demonstrates that the GMD does not perform well with

this kind of variables. It is odd that there are some continuous non-normal distributed variables with high weights. Maybe these continuous variables can adapt and form a joint ellipsoidal distribution more easily than categorical ones, and this way form a more efficient GMD. It is noteworthy that there a lot more weights equal to zero when genetic matching with GAM PS estimates is applied, i.e. there is not the need to resort so heavily to the variables themselves to achieve the best possible balance (e.g. Fluoroscopy Time). This reflects that these GAM PS estimates summarize way better the variables than the LR PS estimates. The variables that are more imbalanced before matching are the ones that are more prone to be weighted, unless these ones are efficiently represented by the respective PS estimate.

Still, genetic matching did not improve the balance as much as expected. This might have happened because the genetic matching needed more iterations run to reach the maximum possible balance. An interesting fact is that genetic matching can overcome the main weaknesses of the logistic regression PS estimates by matching individually on the variables that are not linear with the treatment assignment logit (Fluoroscopy Time and Systolic BP). This is proved when the logistic regression PS estimates based on the Variables Associated with Treatment Assignment almost achieved the same level of balance as the GAM PS estimates based on All Variables (in the genetic matching context). The expectation that the genetic matchings with logistic regression PS estimates would provide roughly the same level of balance between them was defrauded, as well as between the matching with GAM PS estimates. This may have to do with the genetic matching random nature. That is why the weight matrixes are all so different among the genetic matchings.

Overall the best matching achieved was the genetic matching with all variables plus the GAM PS estimates. This led to ATT*=0.000786. The Approach used would not made any difference among the Transradial Approach Group ("Treated Group"), regarding the occurrence of a peri-procedural Stroke/TIA. This result is consistent with the literature (Raposo et al., 2015).

Regarding stratification, the PS estimates used to stratify were the ones that provided best balance across strata. These were estimated from the GAM PS model with All Variables. The balance within each stratum was not brilliant. Although there were no variables with SMD>0.25, there were a lot of variables with SMD>10. Stratification is a coarser way of matching subjects, and because of that it is no surprise that balance within strata is worse (Austin, 2011a).

Stratification ATE estimate is -0.000386 which means if all operated people were to change from Transfemoral to Transradial Approach the population proportion who suffers a peri-procedural Stroke/TIA is 0.000386 less. This difference is not significant.

To perform IPTW the logistic regression PS estimates based on Variables Associated with Treatment Assignment were chosen, because it is not possible to estimate the estimator variance if the PS are estimated by GAMs. IPTW ATE estimate is -0.000155 which means if all operated people were to change from Transfemoral to Transradial Approach the population proportion who suffers a peri-procedural Stroke/TIA is 0.000155 less. This difference is not significant.

The Covariate Adjustment provided roughly the same results as the previous logistic regressions (traditional and Firth's). This indicates that the variables summarized by the PS, in this context, do not provide much information.

Chapter 6

Conclusion

All the procedures applied in this work pointed to the same conclusion. The use of Transradial or Transfemoral Approach did not influence the occurrence of periprocedural Stroke/TIA. This way, the clinically preferential use of Transradial Approach can presumably be performed with little concerns relative to periprocedural neurological complications related to this technique. This result is in accordance with the literature (Hamon et al., 2007; Ratib et al., 2013b; Raposo et al., 2015).

Regarding the methodological issues, it was proved that GAMs provide PS estimates that origin more successful matchings than logistic regression PS estimates, when the variables are not linear with the treatment assignment logit. On the other hand, applying genetic matching can overcome these differences by matching individually on these non-linear variables using the GMD. This conclusions about the methodology reinforce previous results found in the literature that highlight the utility of GAMs in PS estimation and its synergy with genetic matching (Hastie and Tibshirani, 1986; Woo et al., 2008).

Other Work Accomplished

During the course of this work, the author had the opportunity to develop a ShinyR application (web application framework for R). This application "Propensity Scores: An application to Medicine" is the first of its type to analyse in an automated way data through Propensity Score analysis (PS model building, Matching, Genetic Matching, IPTW and Stratification). This application enables the data input and, consequently its intuitive, reliable and time-saving analysis. It can be accessed through this link: $https://luis-garcez-ferreira.shinyapps.io/my\_app\_shiny\_3/$.

Bibliography

Abadie, A. and Imbens, G. W. (2002). Simple and Bias-Corrected Matching Estimators for Average Treatment Effects. *NBER Technical Working Papers*.

Abadie, A. and Imbens, G. W. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267.

Amaral Turkman, M. and Loiola Silva, G. (2000). *Modelos Lineares Generalizados -da teoria à prática-*. Sociedade Portuguesa de Estatística, Lisbon.

Arbogast, P. G. and VanderWeele, T. J. (2013). *Considerations for Statistical Analysis*. Agency for Healthcare Research and Quality (US), Rockville.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27(12).

Austin, P. C. (2011a). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate behavioral research*, 46(3):399–424.

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–61.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, 26(4):734–753.

Bhat, T., Teli, S., Bhat, H., Akhtar, M., Meghani, M., Lafferty, J., and Gala, B. (2012). Access-site complications and their management during transradial cardiac catheterization. *Expert review of cardiovascular therapy*, 10(5):627–34.

Bourassa, M. G. (2005). The history of cardiac catheterization. *The Canadian journal of cardiology*, 21(12):1011–4.

Braitman, L. E. and Rosenbaum, P. R. (2002). Rare outcomes, common treatments: analytic strategies using propensity scores. *Annals of internal medicine*, 137(8):693–5.

Buja, A., Hastie, T., and Tibshirani, R. (1989). *Linear smoothers and additive models*. University of Toronto, Dept. of Statistics, Toronto.

Burzotta, F., Mariani, L., Trani, C., Coluccia, V., Brancati, M. F., and Porto, I. (2013). Management and timing of access-site vascular complications occurring after trans-radial percutaneous coronary procedures. *International journal of cardiology*, 167(5):1973–8.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*.

Cochran, W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*.

D 'agostino, R. B. (1998). Tutorial in Biostatistics Propensity Score Methods for Bias Reduction in the Comparison of a Treatment To a Non-Randomized Control Group. *Statistics in Medicine*, 17:2265–2281.

Dangoisse, V., Guedes, A., Gabriel, L., Jamart, J., Chenu, P., Marchandise, B., and Schroeder, E. (2013). Full conversion from transfemoral to transradial approach for percutaneous coronary interventions results in a similar success rate and a rapid reduction of in-hospital cardiac and vascular major events. *EuroIntervention*, 9(3):345–52.

Deb, S., Austin, P. C., Tu, J. V., Ko, D. T., Mazer, C. D., Kiss, A., and Fremes, S. E. (2015). Methods in Cardiovascular Research: A Review of Propensity-Score Methods and Their Use in Cardiovascular Research. *Canadian Journal of Cardiology*.

Diamond, A. and Sekhon, J. (2012). Genetic Matching for Estimating Causal Effects. *The Review of Economics and Statistics*, 95(July):932–945.

Donnan, G. A., Fisher, M., Macleod, M., and Davis, S. M. (2008). Stroke. *The Lancet*, 371(9624):1612–1623.

Emura, T., Wang, J., and Katsuyama, H. (2008). Assessing the Assumption of Strongly Ignorable Treatment Assignment Under Assumed Causal Models.

Firth, D. (1993). Bias Reduction of Maximum Likelihood. *Biometrika*, 80(1):27–38.

Flury, B. K. and Riedwyl, H. (1986). Standard Distance in Univariate and Multivariate Analysis. *The American Statistician*, 40(3):249.

Friedman, J. H. and Stuetzle, W. (1981). Projection Pursuit Regression. *Journal of the American Statistical Association*, 76(376):817–823.

Gordis, L. (2014). *Epidemiology*. Elsevier Health Sciences, Philadelphia.

Gray, H. (2000). *Anatomy of the Human Body*. Bartebly, New York, 20th edition.

Guo, S. and Fraser, M. W. (2009). *Propensity Score Analysis: Statistical Methods and Applications*. SAGE Publications, Thousand Oaks.

Hamon, M., Gomes, S., Clergeau, M.-R., Fradin, S., Morello, R., and Hamon, M. (2007). Risk of Acute Brain Injury Related to Cerebral Microembolism During Cardiac Catheterization Performed by Right Upper Limb Arterial Access. *American Heart Association Journal*.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1(3):297–310.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Chapman and Hall, London.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419.

Hirano, K. and Imbens, G. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology*, 2(3):259–278.

Huppler Hullsiek, K. and Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics*, 2(4):179–193.

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity : a Review. *The Review of Economics and Statistics*, 86(1):4–29.

Jurga, J., Nyman, J., Tornvall, P., Mannila, M. N., Svenarud, P., van der Linden, J., and Sarkar, N. (2011). Cerebral microembolism during coronary angiography: a randomized comparison between femoral and radial arterial access. *Stroke; a journal of cerebral circulation*, 42(5):1475–7.

Leontief, W. (1947). Introduction to a Theory of the Internal Structure of Functional Relationships. *Econometrica*, 15(4):361.

Lunceford, J. and Davidian, M. (2004). Stratiÿcation and weighting via the propensity score in estimation of causal treatment e ects: a comparative study. *Statistics in Medicine*, 2960(19):2937–2960.

Mazumdar, M. and Glassman, J. R. (2000). Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statistics in medicine*, 19(1):113–32.

Meldrum, M. L. (2000). A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. *Hematology/oncology clinics of North America*, 14(4):745–60, vii.

Mendis; Shanthi, Puska; Pekka, and Norrving; Bo (2011). *Global Atlas on cardiovascular disease prevention and control*. World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization, Geneva.

Mitchell, M. D., Hong, J. A., Lee, B. Y., Umscheid, C. A., Bartsch, S. M., and Don, C. W. (2012). Systematic review and cost-benefit analysis of radial artery access for coronary angiography and intervention. *Circulation. Cardiovascular quality and outcomes*, 5(4):454–62.

Nathan, S. and Rao, S. V. (2012). Radial versus femoral access for percutaneous coronary intervention: implications for vascular complications and bleeding. *Current cardiology reports*, 14(4):502–9.

National Heart, L. and Institute, B. (2016). What Is Coronary Heart Disease?

Normand, S. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., and McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of clinical epidemiology*, 54(4):387–98.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379.

Raposo, L., Madeira, S., Teles, R. C., Santos, M., Gabriel, H. M., Goncalves, P., Brito, J., Leal, S., Almeida, M., and Mendes, M. (2015). Neurologic complications after transradial or transfemoral approach for diagnostic and interventional cardiac catheterization: A propensity score analysis of 16,710 cases from a single centre prospective registry. *Catheterization and Cardiovascular Interventions*, 86(1):61–70.

Ratib, K., Mamas, M., Routledge, H., Ludman, P., Fraser, D., and Nolan, J. (2013a). Influence of access site choice on incidence of neurologic complications after percutaneous coronary intervention. *American heart journal*, 165(3):317–24.

Ratib, K., Mamas, M. A., Routledge, H. C., Ludman, P. F., Fraser, D., and Nolan, J. (2013b). Influence of access site choice on incidence of neurologic complications after percutaneous coronary intervention. *American heart journal*, 165(3):317–24.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *The American Statistician*, 39(1):33–38.

Rubin, D. and Rosenbaum, P. R. (1983). a. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70 SRC -(1):41–55.

Rubin, D. B. (1973). Matching to Remove Bias in Observational Studies. *Biometrics*.

Sacco, R., Benjamin, E., Broderick, J., Dyken, M., Easton, J. D., Feinberg, W. M., Goldstein, L. B., Gorelick, P. B., Howard, G., Kittner, S. J., Manolio, T. A., Whisnant, J. P., and Wolf, P. A. (1997). Risk Factors. *Stroke*, 28(7).

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Wadsworth Cengage Learning, Belmont.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., and Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of clinical epidemiology*, 59(5):437–47.

Topol, E. J. and Teirstein, P. S. (2016). *Textbook of interventional cardiology*. Elsevier, Philadelphia.

Woo, M.-J., Reiter, J. P., and Karr, A. F. (2008). Estimation of propensity scores using generalized additive models. *Statistics in Medicine*, 27(19):3805–3816.

Appendices

Appendix A

Logistic Regression

```
library(ROCR)
library(PredictABEL)
library(mgcv)
library(pROC)

#####DATA#####

data1=read.csv2('cardiologia1e2.csv')
data1=as.data.frame(data1)
names(data1)=c("Age",
               "Male_Gender",
               "BMI",
               "Diabetes_Mellitus",
               "Hypertension",
               "Smoking",
               "Dyslipidemia",
               "MI",
               "PCI",
               "CABG",
               "Stroke_TIA_prev",
               "PAD",
               "Non_CABG_Surgery",
               "Moderate_Severe_CRD",
               "Renal_Transplant",
               "Interventional_Procedure",
               "Acute_Coronary_Syndrome",
               "Aortic_Valvulopathy",
               "Systolic_Blood_Pressure",
               "Diastolic_Blood_Pressure",
               "Coronary_Artery_Disease_Extension",
               "Number_of_Treated_Segments",
               "Fluoroscopy_Time_min",
               "Contrast_Volume",
               "Number_of_Catheters_Used",
```

```
                "Transradial_Approach",   #treatment
                "Stroke_TIA_outcome")     #outcome


###############CHECKING LOGIT LINEARITY###############

par(mfrow=c(2,3))
ylim=c(-2,2)

gam=gam(Stroke_TIA_outcome~s(Age),subset(data1,Age>27&Age<95), family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="Age", ylim=ylim)

gam=gam(Stroke_TIA_outcome~s(BMI),subset(data1,BMI>17&BMI<45), family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="BMI", ylim=ylim)

gam=gam(Stroke_TIA_outcome~s(Systolic_Blood_Pressure),
        subset(data1,Systolic_Blood_Pressure<240), family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="Systolic Blood Pressure", ylim=ylim)

gam=gam(Stroke_TIA_outcome~s(Diastolic_Blood_Pressure),
        subset(data1,Diastolic_Blood_Pressure<100&Diastolic_Blood_Pressure>40),
        family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="Diastolic Blood Pressure", ylim=ylim)

gam=gam(Stroke_TIA_outcome~s(Fluoroscopy_Time_min),
        subset(data1,Fluoroscopy_Time_min<55), family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="Fluoroscopy Time", ylim=ylim)

gam=gam(Stroke_TIA_outcome~s(Contrast_Volume),
        subset(data1,Contrast_Volume<400&Contrast_Volume>10), family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="Contrast Volume",ylim=ylim)


###############MINIMUM p-VALUE APPROACH###############

minp <- function(x, ybin, xcutint) {
  tmp1 <- sapply (sort(unique(xcutint)), function(x0, x, ybin)
  {
    tmp <- chisq.test(1*(x <= x0), ybin)
    tabl <- table(1*(x > x0), ybin)
    rr <- ((tabl[1,1] + 0.5)/(( tabl[1,1] + 0.5) + (tabl[2,1] + 0.5)))/
      ((tabl[1,2] + 0.5)/(( tabl[1,2] + 0.5) + (tabl[2,2] + 0.5)))
    c(x0, tmp$statistic, tmp$p.value, rr)
  }
  , x, ybin)
  tmp1 <- data.frame(t(tmp1))
  tmp1
```

```
}
ybin <- Stroke_TIA_outcome

par(mfrow=c(2,3))

x = Age; xcutint <- 52:68 ;cutpoints=minp(x, ybin, xcutint)[1,]
p_values=minp(x, ybin, xcutint)[3,]
plot(cutpoints, p_values, xlab="cutpoint for Age", ylab="p-value",
     col="red", pch=19, ylim=c(0,1)); abline(h=0.1289, lty=3)
abline(v=65, lty=3)

x = BMI; xcutint = 20:40 ; cutpoints=minp(x, ybin, xcutint)[1,]
p_values=minp(x, ybin, xcutint)[3,]
plot(cutpoints, p_values, xlab="cutpoint for BMI", ylab="p-value",
     col="red", pch=19, ylim=c(0,1)); abline(h=0.3090, lty=3)
abline(v=32, lty=3)

x = Systolic_Blood_Pressure; xcutint=133:155
cutpoints=minp(x, ybin, xcutint)[1,]; p_values=minp(x, ybin, xcutint)[3,]
plot(cutpoints, p_values, xlab="cutpoint for Systolic Blood Pressure",
     ylab="p-value", col="red", pch=19, ylim=c(0,1))
abline(h=0.1022, lty=3); abline(v=147, lty=3)

x = Diastolic_Blood_Pressure; xcutint= 60:100
cutpoints=minp(x, ybin, xcutint)[1,]; p_values=minp(x, ybin, xcutint)[3,]
plot(cutpoints, p_values, xlab="cutpoint for Diastolic Blood Pressure"
     ylab="p-value", col="red", pch=19, ylim=c(0,1));
abline(h=0.0456, lty=3); abline(v=72, lty=3)

x = Fluoroscopy_Time_min; xcutint=5:16
cutpoints=minp(x, ybin, xcutint)[1,]; p_values=minp(x, ybin, xcutint)[3,]
plot(cutpoints, p_values, xlab="cutpoint for Fluoroscopy Time",
     ylab="p-value", col="red", pch=19, ylim=c(0,1))
abline(h=0.16, lty=3); abline(v=10, lty=3)

x = Contrast_Volume; xcutint=80:119
cutpoints=minp(x, ybin, xcutint)[1,]; p_values=minp(x, ybin, xcutint)[3,]
plot(cutpoints, p_values, xlab="cutpoint for Contrast Volume",
     ylab="p-value", col="red", pch=19, ylim=c(0,1))
abline(h=0.37, lty=3); abline(v=92, lty=3)


###############LOGISTIC REGRESSION FINAL###############

Age_65=Age>65
BMI_32=BMI>32
Systolic_Blood_Pressure_147=Systolic_Blood_Pressure>147
```

```
Diastolic_Blood_Pressure_72=Diastolic_Blood_Pressure>72
Fluoroscopy_Time_min_10=Fluoroscopy_Time_min>10

m1=glm(Stroke_TIA_outcome~Transradial_Approach + Dyslipidemia + Age_65 +
         Systolic_Blood_Pressure_72 + Diastolic_Blood_Pressure_147 +
         Acute_Coronary_Syndrome + Fluoroscopy_Time_min_10, family=binomial)
summary(m1)
m2=update(m1, Stroke_TIA_outcome~.-Systolic_Blood_Pressure_147)
summary(m2)
m3=update(m2, Stroke_TIA_outcome~.-Acute_Coronary_Syndrome)
summary(m3)

attach(data1)
model1=glm(Stroke_TIA_outcome~Transradial_Approach+Dyslipidemia + Age_65 +
             Diastolic_Blood_Pressure_72 + Fluoroscopy_Time_min_10,
             family=binomial)
summary(model1)

#subjects used to fit the model
subjects1=which(is.na(Fluoroscopy_Time_min)==FALSE
                &is.na(Diastolic_Blood_Pressure)==FALSE)


###############Calibration Plot###############
detach(data1)

axis=c(0,0.0061)
cOutcome=27
par(mfrow=c(1,2))
plotCalibration(data=data1[subjects1,], cOutcome=27, predRisk=model1$fitted,
                groups=10, rangeaxis=axis)


###############ROC curve###############

plotROC(data=data1[subjects1,], cOutcome=cOutcome,
        predrisk=model1$fitted.values)
roc1 = roc(data1[,cOutcome],model1$fitted)
ci(roc1)
plot(roc1)
```

Appendix B

Firth's Correction

```
library(logistf)

###############UNIVARIATE FIRTH'S LOGISTIC REGRESSION###############

m=logistf(Stroke_TIA_outcome~Renal_Transplant, family=binomial)
summary(m)

m=logistf(Stroke_TIA_outcome~Stroke_TIA_prev, family=binomial)
summary(m)


###############FINAL FIRTH'S LOGISTIC REGRESSION###############

m11=logistf(Stroke_TIA_outcome~Transradial_Approach + Dyslipidemia
            + Age_65 + Systolic_Blood_Pressure_147
            + Diastolic_Blood_Pressure_72
            + Acute_Coronary_Syndrome + Fluoroscopy_Time_min_10
            + Stroke_TIA_prev, family=binomial)

summary(m11)
m22=update(m11, Stroke_TIA_outcome~.-Systolic_Blood_Pressure_147)
summary(m22)
m33=update(m22, Stroke_TIA_outcome~.-Acute_Coronary_Syndrome)
summary(m33)
m44=update(m33, Stroke_TIA_outcome~.-Stroke_TIA_prev)
summary(m44)
model1=m44


################CALIBRATION PLOT################

detach(data1)
axis=c(0,0.0061)
cOutcome=27
```

```
par(mfrow=c(1,2))
plotCalibration(data=data1[subjects1,], cOutcome=27,
                predRisk=model1$predict,
                groups=10, rangeaxis=axis)

################ROC CURVE################

cOutcome=27
plotROC(data=data1[subjects1,], cOutcome=cOutcome,
        predrisk=model1$predict)
roc1 = roc(data1[,cOutcome],model1$fitted)
ci(roc1)
plot(roc1)
```

Appendix C

Propensity Score Matching

```
library("mgcv")
library("ROCR")
library("Matching")
library("memisc")
library("tableone")
library("MKmisc")
library("leaps")

source("functions.R",TRUE)
#contains functions:
#MATCH, MATCH_graph, MATCH_balance, AUCcalib,
#gen_match_model, MATCH_genetic,
#described in Appendix "Functions"


###############TESTING CONFOUNDING###############

data2=cbind(data1,Age_65,BMI_32,Systolic_Blood_Pressure_147,
            Diastolic_Blood_Pressure_72,
            Fluoroscopy_Time_min_10)
attach(data2)

n_var=length(names(data2))-2
uni_out=lapply(names(data2)[c(1:25,28:32)], function(var) {
                    formula     <- as.formula(paste("Stroke_TIA_outcome_~", var))
                    res.logist <- glm(formula, family = binomial)
           c( summary(res.logist)$coef[2,1] , summary(res.logist)$coef[2,4]) }
names(uni_out)=names(data2)[c(1:25,28:32)]
uni_out

uni_exp=lapply(names(data2)[c(1:25,28:32)], function(var) {
                    formula  = as.formula(paste("Transradial_Approach_~", var))
                    res.logist = glm(formula, family = binomial)
                    c( summary(res.logist)$coef[2,1] ,
```

```
                    summary(res.logist)$coef[2,4]) }  )
names(uni_exp)=names(data2)[c(1:25,28:32)]
uni_exp

uni_out_plus_exp=glm(Stroke_TIA_outcome~Transradial_Approach, family=binomial)
coef=uni_out_plus_exp$coef[2]
summary(uni_out_plus_exp)

uni_out_plus_exp_var=lapply(names(data2)[c(1:25,28:32)], function(var) {
  formula= as.formula(paste("Stroke_TIA_outcome ~ Transradial_Approach +",var))
  res.logist= glm(formula, family = binomial)
  summary(res.logist)$coef[2,1]  }  )
names(uni_out_plus_exp_var)=names(data2)[c(1:25,28:32)]
uni_out_plus_exp_var

col_names=c("Outcome~Var Beta",
            "p-value",
            "Exposure~Var Beta",
            "p-value",
            "Outcome~Exposure Beta",
            "Outcome~Exposure+Var (Beta regarding Exposure)",
            "Beta_Variation(%)",
            "OR",
            "Outcome~Exposure+Var (OR regarding Exposure)",
            "OR_Variation(%)")

matrix=matrix(nrow=n_var, ncol=10,
              dimnames=list(names(data2)[c(1:25,28:32)],col_names))
for (i in c(1:30)){ matrix[i,1]=uni_out[[i]][1]
                    matrix[i,2]=uni_out[[i]][2]
                    matrix[i,3]=uni_exp[[i]][1]
                    matrix[i,4]=uni_exp[[i]][2]
                    matrix[i,5]=coef
                    matrix[i,6]=uni_out_plus_exp_var[[i]]
                    matrix[i,7]=(matrix[i,5]-matrix[i,6])/matrix[i,5]*100
                    matrix[i,8]=exp(coef)
                    matrix[i,9]=exp(uni_out_plus_exp_var[[i]])
                    matrix[i,10]=(matrix[i,8]-matrix[i,9])/matrix[i,8]*100}

matrix
matrix1=matrix[which(matrix[,2]<0.25),]
matrix1
matrix2=matrix1[which(matrix1[,4]<0.05),]
matrix2
write.matrix(matrix, file="confounding", sep=",")
```

```
###############BALANCE BEFORE MATCHING###############

tab_baseline = CreateTableOne(vars = names(data2) ,
                              factorVars = names(data2)[c(2,4,5,6,7,8,9,10,
                                                          11,12,13,14,15,16,
                                                          17,18,28,29,
                                                          30,31,32)],
                              strata = "Transradial_Approach" ,
                              data=data2)
print(tab_baseline ,smd=TRUE)
par(mfrow=c(1,1))
index_treated=which(Transradial_Approach==1)
index_control=which(Transradial_Approach==0)
names(data1)

qqplot(Age[index_treated], Age[index_control], ylim=c(0,100), xlim=c(0,100),
       xlab="Age among Transradial Group", ylab="Age among Transfemoral Group",
       main="Before Matching")
abline(a=0,b=1, col=2)

qqplot(Systolic_Blood_Pressure[index_treated],
       Systolic_Blood_Pressure[index_control],
       ylim=c(50,250), xlim=c(50,250),
       xlab="Systolic Blood Pressure among Transradial Group",
       ylab="Systolic Blood Pressure among Transfemoral Group",
       main="Before Matching")
abline(a=0,b=1, col=2)

qqplot(Diastolic_Blood_Pressure[index_treated],
       Diastolic_Blood_Pressure[index_control],
       ylim=c(30,140), xlim=c(30,140),
       xlab="Diastolic Blood Pressure among Transradial Group",
       ylab="Diastolic Blood Pressure among Transfemoral Group",
       main="Before Matching")
abline(a=0,b=1, col=2)

qqplot(Fluoroscopy_Time_min[index_treated],
       Fluoroscopy_Time_min[index_control],
       ylim=c(0,90), xlim=c(0,90),
       xlab="Fluoroscopy Time among Transradial Group",
       ylab="Fluoroscopy Time among Transfemoral Group",,
       main="Before Matching")
abline(a=0,b=1, col=2)


###############PS MODEL (GLM with ALL VARIABLES) AND MATCHING###############
```

```
ps_all_cont=glm(Transradial_Approach~Age+Male_Gender+BMI+Diabetes_Mellitus+
                Hypertension+Smoking+Dyslipidemia+MI+PCI+CABG+Stroke_TIA_prev+
                PAD+Non_CABG_Surgery+Moderate_Severe_CRD+Renal_Transplant+
                Interventional_Procedure+Acute_Coronary_Syndrome+
                Aortic_Valvulopathy+Systolic_Blood_Pressure+
                Diastolic_Blood_Pressure+Coronary_Artery_Disease_Extension+
                Number_of_Treated_Segments+Fluoroscopy_Time_min+
                Contrast_Volume+Number_of_Catheters_Used,
              family=binomial)

#subjects used to fit the PS model
subjects_all=which(is.na(Age+Male_Gender+BMI+Diabetes_Mellitus+
                                Hypertension+Smoking+Dyslipidemia+MI+PCI+
                                CABG+Stroke_TIA_prev+
                                PAD+Non_CABG_Surgery+Moderate_Severe_CRD+
                                Renal_Transplant+Interventional_Procedure+
                                Acute_Coronary_Syndrome+Aortic_Valvulopathy+
                                Systolic_Blood_Pressure+Diastolic_Blood_Pressure+
                                Coronary_Artery_Disease_Extension+
                                Number_of_Treated_Segments+Fluoroscopy_Time_min+
                                Contrast_Volume+Number_of_Catheters_Used)==FALSE)

#evaluate PS model
AUCcalib(ps_all_cont, subjects_all, 26, data1, 1,"All_Variables") #ROC curve
legend(0.4,0.4, legend="_______0.718_\n_(0.709-0.727)", cex=1.3,  bty = "n")
AUCcalib(ps_all_cont, subjects_all, 26, data1, 2) #hosmer
AUCcalib(ps_all_cont, subjects_all, 26, data1, 3, "All_variables") #Hosmer

#matching
match=MATCH(ps_all_cont, subjects_all)
summary(match) #match

#accessing balance after matching
MATCH_balance(subjects_all, data2, match)
MATCH_graph(subjects_all,data2, 1, match)
MATCH_graph(subjects_all,data2, 2, match)
MATCH_graph(subjects_all,data2, 3, match)
MATCH_graph(subjects_all,data2, 4, match)
MATCH_balance(subjects_all, data2, match)
summary(match)

#PS transformation
histogram(log((1-ps_all_cont$fitted)/ps_all_cont$fitted),breaks=100,
          xlab="Transformed_PS_estimates")
histogram(ps_all_cont$fitted,breaks=100,
          xlab="Crude_PS_estimates")
```

```
#genetic matching
X=cbind((data2[,c(1:25)])[subjects_all,],
        log((1-ps_all_cont$fitted)/ps_all_cont$fitted))

ps_all_cont_gen_w =
  gen_match_model( X=X , subjects_all,
                   w=diag(ps_all_cont_gam_gen_w$Weight.matrix))
match=MATCH_genetic(X, subjects_all, ps_all_cont_gen_w )
ps_all_cont_gen=MATCH_balance(subjects_all, data2, match)
MATCH_graph( subjects_all ,data1,1, match)
MATCH_graph( subjects_all ,data1,2, match)
MATCH_graph( subjects_all ,data1,3, match)
MATCH_graph( subjects_all ,data1,4, match)

summary(match) #best matching overall

############PS MODEL (GAM with ALL VARIABLES) AND MATCHING####################

ps_all_cont_gam=gam(Transradial_Approach~s(Age)+Male_Gender+s(BMI)+
                    Diabetes_Mellitus+Hypertension+Smoking+
                    Dyslipidemia+MI+PCI+CABG+Stroke_TIA_prev+
                    PAD+Non_CABG_Surgery+Moderate_Severe_CRD+Renal_Transplant+
                    Interventional_Procedure+Acute_Coronary_Syndrome+
                    Aortic_Valvulopathy+s(Systolic_Blood_Pressure)+
                    s(Diastolic_Blood_Pressure)+
                    Coronary_Artery_Disease_Extension+
                    Number_of_Treated_Segments+
                    s(Fluoroscopy_Time_min)+s(Contrast_Volume)+
                    Number_of_Catheters_Used,
            family=binomial)

#evaluate PS model
AUCcalib(ps_all_cont_gam, subjects_all, 26, data1, 1,"All Variables (GAM)")
legend(0.4,0.4, legend=" 0.750 \n (0.741-0.759)", cex=1.3,  bty = "n")
AUCcalib(ps_all_cont_gam, subjects_all, 26, data1, 2)
AUCcalib(ps_all_cont_gam, subjects_all, 26, data1, 3, "All variables (GAM)")

ylim=c(-2,2)

#matching
match=MATCH(ps_all_cont_gam, subjects_all)

#acessing balance
MATCH_balance(subjects_all, data2, match)
MATCH_graph(subjects_all,data2, 1, match)
MATCH_graph(subjects_all,data2, 2, match)
MATCH_graph(subjects_all,data2, 3, match)
```

```
MATCH_graph(subjects_all,data2, 4, match)

summary(match)


#genetic matching
X=cbind((data2[,c(1:25)])[subjects_all,],
        log((1-ps_all_cont_gam$fitted)/ps_all_cont_gam$fitted))

ps_all_cont_gam_gen_w = gen_match_model( X=X , subjects_all,
                                         w=ps_all_cont_gam_gen_w$Weight.matrix)
match=MATCH_genetic(X, subjects_all, ps_all_cont_gam_gen_w )
ps_all_cont_gen=MATCH_balance(subjects_all, data2, match)
MATCH_graph( subjects_all ,data1,1, match)
MATCH_graph( subjects_all ,data1,2, match)
MATCH_graph( subjects_all ,data1,3, match)
MATCH_graph( subjects_all ,data1,4, match)


############PS MODEL (GLM with Variables Associated with Treatment Assignment)
# AND MATCHING ############
summary(ps_all_cont)
ps_exp_sig1=update(ps_all_cont,~.-Coronary_Artery_Disease_Extension )
summary(ps_exp_sig1)
ps_exp_sig2=update(ps_exp_sig1,~.- Smoking)
summary(ps_exp_sig2)
ps_exp_sig3=update(ps_exp_sig2,~.-Non_CABG_Surgery)
summary(ps_exp_sig3)
ps_exp_sig4=update(ps_exp_sig3,~.-PAD)
summary(ps_exp_sig4)
ps_exp_sig5=update(ps_exp_sig4,~.-Hypertension)
summary(ps_exp_sig5)
ps_exp_sig6=update(ps_exp_sig5,~.-Male_Gender)
summary(ps_exp_sig6)
ps_exp_sig_cont=ps_exp_sig6

subjects_ps_exp_sig_cont=which(is.na(Age + BMI + Diabetes_Mellitus +
                         Dyslipidemia + MI + PCI + CABG + Stroke_TIA_prev +
                         Moderate_Severe_CRD + Renal_Transplant +
                         Interventional_Procedure + Acute_Coronary_Syndrome +
                         Aortic_Valvulopathy + Systolic_Blood_Pressure +
                         Diastolic_Blood_Pressure +
                         Number_of_Treated_Segments +
                         Fluoroscopy_Time_min + Contrast_Volume +
                         Number_of_Catheters_Used)==FALSE)

#PS model evaluation
```

```
AUCcalib(ps_exp_sig_cont, subjects_ps_exp_sig_cont, 26, data1, 1,
         "Variables Assoc. with Treatment Assignment")
legend(0.4,0.4, legend="       0.718 \n (0.708-0.727)", cex=1.3,  bty = "n")
AUCcalib(ps_exp_sig_cont, subjects_ps_exp_sig_cont, 26, data1, 2,
         "Variables Assoc. with Treatment Assignment")
AUCcalib(ps_exp_sig_cont, subjects_ps_exp_sig_cont, 26, data1, 3,
         "Variables Assoc. with Treatment Assignment")

#matching
match=MATCH(ps_exp_sig_cont, subjects_ps_exp_sig_cont)

#Accessing balance
MATCH_balance(subjects_ps_exp_sig_cont, data2, match)
MATCH_graph(subjects_ps_exp_sig_cont,data2, 1, match)
MATCH_graph(subjects_ps_exp_sig_cont,data2, 2, match)
MATCH_graph(subjects_ps_exp_sig_cont,data2, 3, match)
MATCH_graph(subjects_ps_exp_sig_cont,data2, 4, match)

histogram(Fluoroscopy_Time_min[subjects_ps_exp_sig_cont])

histogram(log((1-ps_exp_sig_cont$fitted)/ps_exp_sig_cont$fitted),breaks=100,
          xlab="Transformed PS estimates")
histogram(ps_exp_sig_cont$fitted,breaks=100,
          xlab="Crude PS estimates")

#genetic matching
X=cbind((data2[c(1,3,4,7,8,9,10,11,14,15,16,
                 17,18,19,20,22,23,24,25)])[subjects_ps_exp_sig_cont,] ,
        log((1-ps_exp_sig_cont$fitted)/ps_exp_sig_cont$fitted))

ps_exp_sig_cont_gen_w = gen_match_model( X=X , subjects_ps_exp_sig_cont,
                                         w=ps_exp_sig_cont_gen_w$Weight.matrix)
match=MATCH_genetic(X, subjects_all, ps_exp_sig_cont_gen_w )
ps_all_cont_gen=MATCH_balance(subjects_all, data2, match)
MATCH_graph( subjects_all ,data1,1, match)
MATCH_graph( subjects_all ,data1,2, match)
MATCH_graph( subjects_all ,data1,3, match)
MATCH_graph( subjects_all ,data1,4, match)


############PS MODEL (GAM with Variables Associated with Treatment Assignment)
# AND MATCHING ############

ps_exp_sig_cont_gam=gam(Transradial_Approach~s(Age)+s(BMI)+Diabetes_Mellitus+
                        Dyslipidemia+MI+PCI+CABG+Stroke_TIA_prev+
                        Moderate_Severe_CRD+Renal_Transplant+
                        Interventional_Procedure+
```

```
                                Acute_Coronary_Syndrome+Aortic_Valvulopathy+
                                s(Systolic_Blood_Pressure)+
                                s(Diastolic_Blood_Pressure)+
                                Number_of_Treated_Segments+
                                s(Fluoroscopy_Time_min)+s(Contrast_Volume)+
                                Number_of_Catheters_Used,
                              family=binomial)

#PS model evaluation
AUCcalib(ps_exp_sig_cont_gam, subjects_ps_exp_sig_cont,
         26, data1, 1,"Variables Assoc. with Treatment Assignment \n (GAM)")
legend(0.4,0.4, legend=" 0.750 \n (0.741-0.758)", cex=1.3,  bty = "n")

AUCcalib(ps_exp_sig_cont_gam, subjects_ps_exp_sig_cont,
         26, data1, 2,"Variables Assoc. with Treatment Assignment \n (GAM)")
AUCcalib(ps_exp_sig_cont_gam, subjects_ps_exp_sig_cont,
         26, data1, 3,"Variables Assoc. with Treatment Assignment \n (GAM)")

#matching
match=MATCH(ps_exp_sig_cont_gam, subjects_ps_exp_sig_cont)

#Accessing balance
MATCH_balance(subjects_ps_exp_sig_cont, data2, match)
MATCH_graph(subjects_ps_exp_sig_cont,data2, 1, match)
MATCH_graph(subjects_ps_exp_sig_cont,data2, 2, match)
MATCH_graph(subjects_ps_exp_sig_cont,data2, 3, match)
MATCH_graph(subjects_ps_exp_sig_cont,data2, 4, match)

summary(match)


#genetic matching
X=cbind((data2[c(1,3,4,7,8,9,10,11,14,15,16,
                 17,18,19,20,22,23,24,25)])[subjects_ps_exp_sig_cont,] ,
        log((1-ps_exp_sig_cont_gam$fitted)/ps_exp_sig_cont_gam$fitted))


ps_exp_sig_cont_gam_gen_w =
  gen_match_model( X=X , subjects_ps_exp_sig_cont,
                          w=ps_exp_sig_cont_gam_gen_w$Weight.matrix)

match=MATCH_genetic(X, subjects_all, ps_exp_sig_cont_gam_gen_w )
ps_all_cont_gen=MATCH_balance(subjects_all, data2, match)
MATCH_graph( subjects_all ,data1,1, match)
MATCH_graph( subjects_all ,data1,2, match)
MATCH_graph( subjects_all ,data1,3, match)
MATCH_graph( subjects_all ,data1,4, match)
```

```
###############CHECKING LOGIT LINEARITY REGARDING TREATMENT###############

par(mfrow=c(2,3))
ylim=c(-2,2)

gam=gam(Transradial_Approach~s(Age),subset(data1,Age>27&Age<95),
        family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="Age", ylim=ylim)

gam=gam(Transradial_Approach~s(BMI),subset(data1,BMI>17&BMI<45),
        family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="BMI", ylim=ylim)

gam=gam(Transradial_Approach~s(Systolic_Blood_Pressure),
        subset(data1,Systolic_Blood_Pressure<240), family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="Systolic Blood Pressure", ylim=ylim)

gam=gam(Transradial_Approach~s(Diastolic_Blood_Pressure),
        subset(data1,Diastolic_Blood_Pressure<100&Diastolic_Blood_Pressure>40),
plot(gam, se=T, ylab="f(x)", xlab="Diastolic Blood Pressure", ylim=ylim)

gam=gam(Transradial_Approach~s(Fluoroscopy_Time_min),
        subset(data1,Fluoroscopy_Time_min<55),
        family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="Fluoroscopy Time", ylim=ylim)

gam=gam(Transradial_Approach~s(Contrast_Volume),
        subset(data1,Contrast_Volume<400&Contrast_Volume>10),
        family=binomial)
plot(gam, se=T, ylab="f(x)", xlab="Contrast Volume",ylim=ylim)
```

Appendix D

Stratification

```
library(mgcv)
library("vioplot")
library("tableone")

source("functions.R",TRUE)
#contains functions:
#STRATA, STRATA_balance, STRATA_graph
#described in Appendix "Functions"


################PS MODEL FOR STRATIFICATION###############

#PS model
attach(data1)
ps_all_cont_gam=gam(Transradial_Approach~s(Age)+Male_Gender+s(BMI)+
                      Diabetes_Mellitus+Hypertension+Smoking+Dyslipidemia+
                      MI+PCI+CABG+Stroke_TIA_prev+PAD+Non_CABG_Surgery+
                      Moderate_Severe_CRD+Renal_Transplant+
                      Interventional_Procedure+Acute_Coronary_Syndrome+
                      Aortic_Valvulopathy+s(Systolic_Blood_Pressure)+
                      s(Diastolic_Blood_Pressure)+
                      Coronary_Artery_Disease_Extension+
                      Number_of_Treated_Segments+s(Fluoroscopy_Time_min)+
                      s(Contrast_Volume)+Number_of_Catheters_Used,
                    family=binomial)

subjects_all=which(is.na(Age+Male_Gender+BMI+Diabetes_Mellitus+
                           Hypertension+Smoking+Dyslipidemia+MI+PCI+CABG+
                           Stroke_TIA_prev+PAD+Non_CABG_Surgery+
                           Moderate_Severe_CRD+Renal_Transplant+
                           Interventional_Procedure+Acute_Coronary_Syndrome+
                           Aortic_Valvulopathy+Systolic_Blood_Pressure+
                           Diastolic_Blood_Pressure+
```

```
                                Coronary_Artery_Disease_Extension+
                                Number_of_Treated_Segments+
                                Fluoroscopy_Time_min+Contrast_Volume+
                                Number_of_Catheters_Used)==FALSE)


model=ps_all_cont_gam
subjects=subjects_all

#attribute subjects to strata
strata=STRATA(data1, subjects, model)


###############BALANCE ACROSS STRATA###############

#SMD tables to each stratum
write.csv(STRATA_balance(data1, subjects, stratum1=1, strata),
          "stratum1_balance")
write.csv(STRATA_balance(data1, subjects, stratum1=2, strata),
          "stratum2_balance")
write.csv(STRATA_balance(data1, subjects, stratum1=3, strata),
          "stratum3_balance")
write.csv(STRATA_balance(data1, subjects, stratum1=4, strata),
          "stratum4_balance")
write.csv(STRATA_balance(data1, subjects, stratum1=5, strata),
          "stratum5_balance")

sum(strata==1);sum(strata==2);sum(strata==3);sum(strata==4);sum(strata==5);

par(mfrow=c(4,1))
par(mfrow=c(1,1))
strata=STRATA(data1, subjects, model)

#qq-plots for each strata and each variable (for stratum 1)
STRATA_graph(data1, subjects, stratum1=1, strata, graph=1)
STRATA_graph(data1, subjects, stratum1=1, strata, graph=2)
STRATA_graph(data1, subjects, stratum1=1, strata, graph=3)
STRATA_graph(data1, subjects, stratum1=1, strata, graph=4)


###############STRATIFICATION REPRESENTATION###############

ps=ps_all_cont_gam$fitted
plot(ps[which(treated==1)] ~ c(1:length(ps[which(treated==1)])),
     xlim=c(0,length(subjects)),
     ylim=c(0,1),
     col=4,
```

```
      xlab="Subject Index",
      ylab="PS Estimates")
points(ps[which(treated==0)]~c((length(ps[which(treated==1)])+1):
                                length(subjects)), col=3)
quantiles=quantile(ps, c(0,0.2,0.4,0.6,0.8,1))
abline(h=quantiles, col=2, lwd=2)
legend("topright",c("Transradial","Transfemoral"),
       col=c(4,3), pch=1, cex=0.8, pt.cex=1, bty="o")


#vioplot
vioplot(ps[which(treated==1)],ps[which(treated==0)], col=7,
        names=c("Transradial Group","Transfemoral Group"),
        #ylab="PS Estimates",
        #main="Common Support",
        ylim=c(0,1))
title(ylab="PS Estimates")
abline(h=quantiles, col=2, lwd=2)


################ATE ESTIMATION###############

outcome=Stroke_TIA_outcome[subjects]
treated=Transradial_Approach[subjects]

#stratum specifics ATE
ate1=mean(outcome[which(treated==1&strata==1)])-
  mean(outcome[which(treated==0&strata==1)])
ate2=mean(outcome[which(treated==1&strata==2)])-
  mean(outcome[which(treated==0&strata==2)])
ate3=mean(outcome[which(treated==1&strata==3)])-
  mean(outcome[which(treated==0&strata==3)])
ate4=mean(outcome[which(treated==1&strata==4)])-
  mean(outcome[which(treated==0&strata==4)])
ate5=mean(outcome[which(treated==1&strata==5)])-
  mean(outcome[which(treated==0&strata==5)])

ATE=mean(c(ate1,ate2,ate3,ate4,ate5))


################VARIANCE###############

#stratum specific variance
sd_stratum=function(treated1, outcome1, strata1, stratum1){

  subjects_stratum=which(strata1==stratum1)
  subjects_stratum_1=which(treated1==1&strata1==stratum1)
```

```
  subjects_stratum_0=which(treated1==0&strata1==stratum1)
  n_stratum_treated=length(subjects_stratum_1)
  n_stratum_control=length(subjects_stratum_0)

  y1j = mean(outcome1[subjects_stratum_1])
  y0j = mean(outcome1[subjects_stratum_0])
  var1=(1/n_stratum_treated)*
    sum(((treated1[subjects_stratum]*outcome1[subjects_stratum])-y1j)^2)
  var0=(1/n_stratum_control)*
    sum((((1-treated1[subjects_stratum])*outcome1[subjects_stratum])-y0j)^2)

  var_stratum=var1/n_stratum_treated + var0/n_stratum_control

  return(var_stratum)
}

#overall variance
var_ATE=function(n_strata1, vars1){

  if (n_strata1==length(vars1)){return(1/(n_strata1^2)*sum(vars1))}

}

vars=c(sd_stratum(treated, outcome, strata, 1),
       sd_stratum(treated, outcome, strata, 2),
       sd_stratum(treated, outcome, strata, 3),
       sd_stratum(treated, outcome, strata, 4),
       sd_stratum(treated, outcome, strata, 5))
varATE=var_ATE(5,vars)
sdATE=sqrt(varATE)


################ATE ESTIMATION AND SIGNIFICANCE###############

#p-value
2*(1-pnorm(abs(ATE/sdATE)))


#estimation summary
matrix(c(ate1,ate2,ate3,ate4,ate5, sqrt(vars[1]), sqrt(vars[2]),
         sqrt(vars[3]), sqrt(vars[4]), sqrt(vars[5])),5,2,
       dimnames=list(c("1","2","3","4","5"),c("ATE","SD")))

matrix(c( ATE, sdATE, ATE/sdATE, pnorm(ATE/sdATE)),4,1,
       dimnames=list(c("Overall_ATE","sd_ATE","ATE/sdATE","p-value"),
                     c("OVERALL")))
```

Appendix E

IPTW

```
source("functions.R", TRUE)
#contains functions:
#IPTW_ATE, H_beta1, E_betabeta_inverse1, IPTW_i_variance1, IPTW_total
#described in Appendix "Functions"


#ATE estimation
ate_iptw=IPTW_ATE(data1, subjects_ps_exp_sig, ps_exp_sig_cont)

#Variance
H_beta = H_beta1(data1, subjects_ps_exp_sig, ps_exp_sig_cont)
E_betabeta_inverse=E_betabeta_inverse1(data1, subjects_ps_exp_sig,
                                       ps_exp_sig_cont)
IPTW_i_variance = IPTW_i_variance1(ps_exp_sig_cont, ate_iptw,
                                   data1, subjects_ps_exp_sig ,4,
                                   H_beta, E_betabeta_inverse)
var_IPTW=IPTW_total(data1,subjects_ps_exp_sig, ps_exp_sig_cont,
                    ate_iptw, H_beta, E_betabeta_inverse)
sqrt(var_IPTW)

#p-value
ate_iptw/sqrt(var_IPTW)
2*pnorm(ate_iptw/sqrt(var_IPTW))
```

Appendix F

Functions

```
###############MATCHING###############

#normal matching
MATCH=function(model, subjects ){
  set.seed(1)
  ps_matching = Match(Y=Stroke_TIA_outcome[subjects],
                      Tr=Transradial_Approach[subjects],
                      X= log((1-model$fitted)/model$fitted),
                      ties=FALSE,
                      estimand="ATT",
                      caliper=0.20,
                      replace=TRUE,
                      distance.tolerance = 1e-50
  )
  ps_matching
}

#weight matrix used in genetic matching
gen_match_model=function(X, subjects){

  GenMatch(Transradial_Approach[subjects], X, estimand="ATT", M=1,
           pop.size = 500, fit.func="pvals" , ties=FALSE) #KS test

}

#genetic matching
MATCH_genetic=function(X, subjects ,w){
  set.seed(1)
  ps_matching = Match(Y=Stroke_TIA_outcome[subjects],
                      Tr=Transradial_Approach[subjects], X= X,
                      ties=FALSE,
                      estimand="ATT",
                      replace=TRUE,
                      Weight.matrix=w
```

```
  )
  ps_matching
}


#matching balance (qq-plots)
MATCH_graph=function( subjects ,data_complete, graph, ps_matching){


  data_complete=data_complete[subjects ,]
  data=data_complete

  index_treated=ps_matching$index.treated
  index_control=ps_matching$index.control

  if (graph==1){
  qqplot(data$Age[index_treated],
         data$Age[index_control],
         ylim=c(0,100), xlim=c(0,100),
         xlab="Age among Transradial Group",
         ylab="Age among Transfemoral Group",
         main="After Matching")
  abline(a=0,b=1, col=2)}
  if (graph==2){
  qqplot(data$Systolic_Blood_Pressure[index_treated],
         data$Systolic_Blood_Pressure[index_control],
         ylim=c(50,250), xlim=c(50,250),
         xlab="Systolic Blood Pressure among Transradial Group",
         ylab="Systolic Blood Pressure among Transfemoral Group",
         main="After Matching")
  abline(a=0,b=1, col=2)}
  if (graph==3){
    qqplot(data$Diastolic_Blood_Pressure[index_treated],
           data$Diastolic_Blood_Pressure[index_control],
         ylim=c(30,140), xlim=c(30,140),
         xlab="Diastolic Blood Pressure among Transradial Group",
         ylab="Diastolic Blood Pressure among Transfemoral Group",
         main="After Matching")
  abline(a=0,b=1, col=2)}
  if (graph==4){qqplot(data$Fluoroscopy_Time_min[index_treated],
                       data$Fluoroscopy_Time_min[index_control],
         ylim=c(0,90), xlim=c(0,90),
         xlab="Fluoroscopy Time among Transradial Group",
         ylab="Fluoroscopy Time among Transfemoral Group",
         main="After Matching")
  abline(a=0,b=1, col=2)}
}
```

```
#matching balance  (SMD table)
MATCH_balance=function(subjects,data_complete, ps_matching){

  data_complete=data_complete[subjects,]
  data=data_complete[c(ps_matching$index.treated,ps_matching$index.control),]

  tab_baseline_ps_matching =
    CreateTableOne(vars = names(data_complete) ,
                   factorVars = names(data_complete)[c(2,4,5,6,7,8,9,
                                                       10,11,12,13,14,
                                                       15,16,17,18,28,
                                                       29,30,31,32)],
                   strata ="Transradial_Approach" ,
                   data=data)
  return(print(tab_baseline_ps_matching, smd=TRUE))
}


###############MODEL EVALUATION###############

#AUC and Calibration plot model
AUCcalib=function(model1, subjects1, outcome1, data11, option1, title){

  if (option1==1){
    plotROC(data=data11[subjects1,], cOutcome=outcome1,
            predrisk=model1$fitted.values, plottitle=title)
  }

  if(option1==2){
    return(HLgof.test(model1$fitted,
                      data1[,outcome1][subjects1]))
  }

  if(option1==3){
    plotCalibration(data=data11[subjects1,],
                    cOutcome=outcome1, predRisk=model1$fitted,
                    groups=10, rangeaxis=c(0,1), plottitle=title)}
}


###############STRATIFICATION###############


#strata classification
STRATA=function(data11, subjects1, model1){
```

```
  data11=data11[subjects,]
  treated=Transradial_Approach[subjects]
  ps=model1$fitted
  quantiles=quantile(ps, probs=c(0,0.20,0.40,0.60,0.80,1))
  if (length(ps)==dim(data11)[1]){
  strata=NULL
  for (i in 1:length(subjects1))
          {if (ps[i]<quantiles[2]){strata[i]=1}
          if (ps[i]>quantiles[2]&ps[i]<quantiles[3]){strata[i]=2}
          if (ps[i]>quantiles[3]&ps[i]<quantiles[4]){strata[i]=3}
          if (ps[i]>quantiles[4]&ps[i]<quantiles[5]){strata[i]=4}
          if (ps[i]>quantiles[5]){strata[i]=5}
  }}
  return(strata)
}


#balance across strata (qq-plots)
STRATA_graph=function(data_complete1, subjects1, stratum1, strata1, graph){
  data=data_complete1[subjects1,]
  treated1=data$Transradial_Approach

  index_treated=which(strata1==stratum1&treated1==1)
  index_control=which(strata1==stratum1&treated1==0)

  if (graph==1){
    qqplot(data$Age[index_treated], data$Age[index_control], ylim=c(0,100),
           xlim=c(0,100),
           xlab="Age among Transradial Group",
           ylab="Age among Transfemoral Group",
           main=paste("Stratum", stratum1, sep=" "))
    abline(a=0,b=1, col=2)}
  if (graph==2){
    qqplot(data$Systolic_Blood_Pressure[index_treated],
           data$Systolic_Blood_Pressure[index_control],
           ylim=c(50,250), xlim=c(50,250),
           xlab="Systolic Blood Pressure among Transradial Group",
           ylab="Systolic Blood Pressure among Transfemoral Group",
           main=paste("Stratum", stratum1, sep=" "))
    abline(a=0,b=1, col=2)}
  if (graph==3){
    qqplot(data$Diastolic_Blood_Pressure[index_treated],
           data$Diastolic_Blood_Pressure[index_control],
           ylim=c(30,140), xlim=c(30,140),
           xlab="Diastolic Blood Pressure among Transradial Group",
           ylab="Diastolic Blood Pressure among Transfemoral Group",
```

```
            main=paste("Stratum", stratum1, sep=" "))
    abline(a=0,b=1, col=2)}
  if (graph==4){qqplot(data$Fluoroscopy_Time_min[index_treated],
                       data$Fluoroscopy_Time_min[index_control],
                       ylim=c(0,90), xlim=c(0,90),
                       xlab="Fluoroscopy Time among Transradial Group",
                       ylab="Fluoroscopy Time among Transfemoral Group",
                       main=paste("Stratum", stratum1, sep=" "))
                abline(a=0,b=1, col=2)}

}


#balance across strata (SMD tables)
STRATA_balance=function(data_complete, subjects1, stratum1, strata1){

  data_complete=data_complete[subjects,]
  data=data_complete[which(stratum1==strata1),]

  tab_baseline_ps_matching =
    CreateTableOne(vars = names(data_complete),
                   factorVars = names(data_complete)[c(2,4,5,6,7,8,9,10,11,
                                                       12,13,14,15,16,17,18
                   strata ="Transradial_Approach", data=data)
  return(print(tab_baseline_ps_matching, smd=TRUE))
}


###############IPTW###############


#IPTW variance estimation
H_beta1 = function(data11, subjects1, model11){

  data66 = cbind(rep(1,length(subjects1)),
                 data11[subjects1,][,c(1,3,4,7,8,9,10,11,14,15,16,
                                       17,18,19,20,22,23,24,25)])
  Y=Stroke_TIA_outcome[subjects1]
  Z=Transradial_Approach[subjects1]
  e=model11$fitted
  n = dim(data66)[1]
  W = t(data66)

  sum=0
  for (i in 1:n){
  sum = sum + (Z[i]*Y[i]*(1-e[i])/e[i] + (1-Z[i])*Y[i]*e[i]/(1-e[i])) * W[,i]
  }
```

```
  return( as.matrix((1/n)*sum)   )

}

E_betabeta_inverse1=function(data11, subjects1, model11){

  data66 = cbind(rep(1,length(subjects1)),
                 data11[subjects1,][,c(1,3,4,7,8,9,10,11,14,15,16,
                                       17,18,19,20,22,23,24,25)])
  e=model11$fitted
  n = dim(data66)[1]
  W = t(data66)

  sum=0
  for (i in (1:n)){

    sum = sum + e[i] * (1-e[i] ) * W[,i] %*% t(W[,i])
  }
  return(as.matrix(1/n*sum))
}


IPTW_i_variance1=function(model11, iptw_estimate1, data11,
                          subjects1 ,index1, H_beta11, E_betabeta_inverse1){

  data66 = cbind(rep(1,length(subjects1)),
                 data11[subjects1,][,c(1,3,4,7,8,9,10,11,14,15,16,
                                       17,18,19,20,22,23,24,25)])
  Y=Stroke_TIA_outcome[subjects1]
  Z=Transradial_Approach[subjects1]
  e=model11$fitted
  n = dim(data66)[1]
  W = t(data66)

  part1 = Z[index1]*Y[index1]/e[index1]

  part2 = (1-Z[index1])*Y[index1]/(1-e[index1])

  part3 = iptw_estimate1


  part4 = (Z[index1] - e[index1])%*%
    (t(H_beta11)%*%E_betabeta_inverse1%*%W[,index1])

  return(part1-part2-part3-part4)
}
```

```
IPTW_total=function(data11, subjects1, model11,
                    iptw_estimate1, H_beta11, E_betabeta_inverse1){

  data66 =cbind(rep(1,length(subjects1)),
                data11[subjects1,][,c(1,3,4,7,8,9,10,11,14,15,16,
                                      17,18,19,20,22,23,24,25)])
  n = dim(data66)[1]

  sum=0
  for (i in 1:n){sum=sum+(IPTW_i_variance1(model11, iptw_estimate1,
                                           data11, subjects1 ,i,
                                           H_beta11, E_betabeta_inverse1))}

  return((1/(n^2))*sum^2) #final IPTW estimation variance

}


#IPTW ATE estimation
IPTW_ATE=function(data11, subjects1, model11){

  data66 = data11[subjects1,][,1:25]
  Y=Stroke_TIA_outcome[subjects1]
  Z=Transradial_Approach[subjects1]
  e=model11$fitted
  n = dim(data66)[1]

  ATE= (1/n)*sum(Z*Y/e) - (1/n)*sum(((1-Z)*Y)/(1-e))
  return(c(ATE))

}
```